University of Colorado Boulder

UNIVERSITY OF COLORADO BOULDER

Evaluating the Effectiveness of Using Weather Conditions as a Predictor for Occurrence and Duration of Flight Delays

---

Student :
Ella Bronaugh
Mayur Dalvi
Reza Naiman
Tristan Levy-Park

Teacher :
Dr. Kris Pruitt

Domain Experts :
Dr. Edward Ochoa
Jacob Crampton
Leon Shen
Logan Gage

December 16, 2024

Contents

1 Introduction

Flight delays due to adverse weather conditions pose significant challenges to the aviation industry. Researchers Christopher J. Goodman and Jennifer D. Small Griswold found that extreme weather events were responsible for 32.6% of the total delay minutes recorded in the National Airspace System (NAS) from 2003 to 2015, with severe weather causing up to 82% of delay minutes in some instances [3]. Such delays have an impact on the environment as well as economic performance in the airline industry. [6, 7, 8]. These environmental impacts include increasing fuel consumption and emissions. For example, when aircraft are delayed, they often remain idling on the runway or need to reroute, leading to additional greenhouse gas emissions and contributing to air pollution. The annual economic impact of airline delays was estimated by one study to be $31.2 billion in 2010 and $40.2 billion in other estimates [1]. In addition, climate change and air transportation have a reciprocal relationship: aircraft emissions contribute to anthropogenic climate change while atmospheric changes directly impact operations in the airline industry [5]. Not only this–delays due to inclement weather are increasing over time. In comparing the impact of a winter storm in December 2021 versus December 2022, aggregated total passenger "dwell time" in airports saw an increase of approximately 12 million hours [9]. There are several factors contributing to increasing frequency of weather delays and the economic and environmental impacts are substantial; thus, the ability to accurately predict the occurrence and duration of flight delays and manage these disruptions is increasingly crucial. This study aims to answer the following question: are weather predictors effective at accurately classifying delayed flights?

Previous studies have highlighted the interaction between meteorological conditions and aviation operations, employing various Machine Learning (ML) and data-driven approaches to forecast delays [2, 3, 4]. These findings emphasized the importance of understanding weather patterns and airport-specific vulnerabilities to optimize flight schedules and improve operational efficiency. A group of researchers, Kerim Kiliç and Jose M. Sallan, examined arrival delays across the United States airport network using a variety of models [4]. Their analysis, based on 2017 flight and weather data, found the Gradient Boosting Machine (GBM) model to be the most effective. Although their study covered a larger geographic area, it mainly focused on classifying delays and faced challenges with imbalanced data and limited real-time data use. In a study conducted in 2024, Seongeun Kim and Eunil Park expanded the scope by applying a wider suite of ML models to predict departure delays at three major international airports [2]. Their models achieved high predictive accuracy, with rates of 74.9% for Incheon (ICN), 85.2% for John F. Kennedy (JFK), and 78.5% for Chicago Midway (MDW) in 2-hour forecasts. Although their study demonstrated the potential of ML models in long-term delay predictions, it was limited by its focus on individual airports and a reliance on historical datasets from 2011 to 2021, which may not fully capture future or emerging trends in weather patterns. In a similar study, Sun Choi and others also trained ML models with the goal of classifying whether a flight was delayed or on-time due to weather conditions. Their highest accuracy percentage was 80.36% using the Random Forest (RF) classifier [10]. Though several studies have been conducted in which delay classification prediction was explored, the scope of the datasets has been limited.

Building on these foundational studies, our research aims to address the limitations of previous work by integrating recent, high-frequency data across a more diverse range of U.S. airports. Unlike previous studies that focused on single airports or had limited data, we used advanced model selection and real-time data integration to improve prediction accuracy and generalizability. We compared the accuracy of classification models in three scenarios: (1) when weather conditions aren't used as an independent variable, (2) when they're used in conjunction with flight data as independent variables, and (3) when weather conditions are used as an independent variable without flight data. By using this approach, we aim to determine if weather conditions are an effective predictor for precise models with a high level of accuracy in predicting flight delays. In developing models that account for the dynamic interactions between weather conditions and flight delays, we seek to provide stakeholders with actionable insights to reduce delay-related costs, environmental impacts, and improve overall passenger satisfaction. This research will not only advance the current state of delay prediction, but also contribute to the broader field of transportation analytics, offering scalable solutions to mitigate weather-related disruptions in various modes of transportation. We believe that by using our diverse high-frequency dataset this study will find that weather conditions are an invaluable explanatory variable when it comes to classifying flight delays.

2 Sources

Having a valid and reliable source to find data for a ML project is essential. One of our main goals for this project is to determine the effectiveness of weather data as a predictor for flight delays. To carefully find the relationship between weather and flight delays, we needed a dataset that contained both the weather data and the flight data. However, after exploring online resources for such data, we found that this kind of dataset does not exist. The subsequent alternative to finding the data was to find one source for weather data and one for flight data, and then combine them.

We found the Department of Transportation (DOT) a credible source for flight data. The website allowed us to extract many important variables related to flight information such as flight date, time, airport identification number, departure time, and the number of minutes the flight is delayed (difference in minutes between scheduled and actual departure). The predictors from the flight data we used include variables such as month and year, origin and destination, time of day, airline, and airport. We created a binary column that indicates if a flight has been delayed for 5 minutes or longer which we'll use as the classification response variable. The column containing the number of minutes the flight has been delayed will be used as the response variable for our regression model.

Although numerous sources exist to find weather data such as Weather Underground and National Oceanic and Atmospheric Administration (NOAA), we found Iowa Environmental Mesonet (IEM) a reliable source for collecting the weather data. We found this source credible for the following two reasons. First, IEM solely focuses on airport weather data, not only in the US but also at airports around the world, which can inform future work by applying the findings of this paper to global datasets. Secondly, IEM extracts the data through the Automated Surface Observing System and according to the IEM website, "ASOS networks are nationally monitored for quality 24 hours per day".

This adds another level of confidence and certainty to the validity and reliability of the data. The predictors from the weather data used include variables such as air temperature, dew point temperature, wind speed, wind direction, visibility, and pressure. We included weather and flight data of all the airports from Georgia and Illinois between the years of 2014 and 2024.

After conducting a comprehensive literature review, we identified the key concerns and how different researchers approached weather-related flight delays. Almost all of the literature reviews have been published in the past 5 years which is very relevant to our research. The oldest one goes back to July 2007, which is still relevant today, but this paper only talks about the environmental impacts of flight delays. The studies that we have used as our sources have been published in credible journals in the fields of aviation, aerospace, and meteorology which are all essential fields of studies to understand the complex relationship between weather and flight delays. Not only that, most of our sources have been published by the most credible scientific publishers such as Springer and Cambridge University Press. These interdisciplinary publications provide us with a strong foundation on past work completed in this field and ensure the reliability and validity of our sources.

3 Exploration

This study involved collecting and integrating two primary datasets: weather data and flight data, spanning a decade and covering multiple airports across two states. The data collection and cleaning processes were crucial in ensuring the datasets were accurate, aligned, and ready for analysis.

3.1 Data Collection

Weather Data:

The weather data was sourced from the Automated Surface Observing Systems (ASOS) and Automated Weather Observing Systems (AWOS), accessible via the Iowa Environmental Mesonet (IEM) ASOS Network. This dataset includes METAR-format observations from airport sensors worldwide.

Originally recorded at one-hour intervals, the data was accessed through a custom API request script. Records with missing or invalid temperature data were excluded, and to enable finer analysis, linear interpolation was applied to expand the data to 15-minute intervals.

Flight Data:

Flight data was manually collected from the Bureau of Transportation Statistics (BTS), Department of Transportation, using their online platform. Data for each month was downloaded in ZIP format, spanning the same two states and ten years as the weather data. Then monthly files were decompressed and concatenated to create a comprehensive dataset.

3.2 Data Cleaning and Integration

Alignment by Time and Location:

The two datasets were merged based on datetime and airport station identifiers, linking flight records with corresponding weather data. The one-hour weather data intervals were interpolated to 15-minute intervals for precise temporal alignment. Additionally, the scheduled departure time in the flight data was rounded to the nearest 15 minutes to facilitate a left join with the weather data.

Handling Missing Data:

Missing values in weather data were addressed through forward-filling techniques, ensuring that essential values were available for each interval. Flight data underwent similar treatment, with forward-fill and linear interpolation applied to maintain consistency and prevent data gaps.

This systematic data collection, cleaning, and integration ensured robust alignment and quality, forming a reliable foundation for the subsequent analyses and model development. Table 1 presents the final set of variables included in our final dataset.

Feature	Data Type	Description
Year	int64	Year of the flight
Quarter	int64	Quarter of the year
Month	int64	Month of the flight
Day_of_Month	float64	Day of the month
Day_of_Week	float64	Day of the week
Operating_Carrier_Code	object	Code of the operating airline
Tail_Number	object	Aircraft tail number
Origin_Airport_ID	float64	ID of the origin airport
Origin_Airport_Code	object	IATA code of the origin airport
Origin_State_Name	object	State name of the origin airport
Destination_Airport_Code	object	IATA code of the destination airport
Destination_State_Name	object	State name of the destination airport
Scheduled_Departure_Time	float64	Scheduled departure time in minutes
Departure_Delay_Minutes	float64	Departure delay in minutes
Taxi_Out_Time_Minutes	float64	Taxi-out time in minutes
Flight_Distance_Miles	float64	Distance of the flight in miles
Departure_Datetime	object	Exact departure datetime
Scheduled_Departure_Time_Minutes	float64	Scheduled departure time in minutes
Air_Temperature_Fahrenheit	float64	Air temperature at the origin airport
Dew_Point_Temperature_Fahrenheit	float64	Dew point temperature at the origin airport
Relative_Humidity_Percent	float64	Relative humidity at the origin airport
Wind_Direction_Degrees	float64	Wind direction in degrees at the origin airport
Wind_Speed_Knots	float64	Wind speed in knots at the origin airport
Hourly_Precipitation_Inches	float64	Hourly precipitation in inches
Pressure_Altimeter_Inches	float64	Altimeter pressure in inches
Sea_Level_Pressure_Millibar	float64	Sea level pressure in millibars
Visibility_Miles	float64	Visibility in miles
Sky_Cover_Level_1	object	Sky cover description
Sky_Level_1_Altitude_Feet	float64	Sky cover altitude in feet
Apparent_Temperature_Fahrenheit	float64	Apparent temperature at the origin airport
Target	float64	Target variable for prediction

Table 1: Final Feature Set for Flight Delay Prediction Project

3.3 Data Balancing

Initial distribution:

After our data collection and cleaning process, the distribution of non-delayed (0) versus delayed (1) flights was as follows:

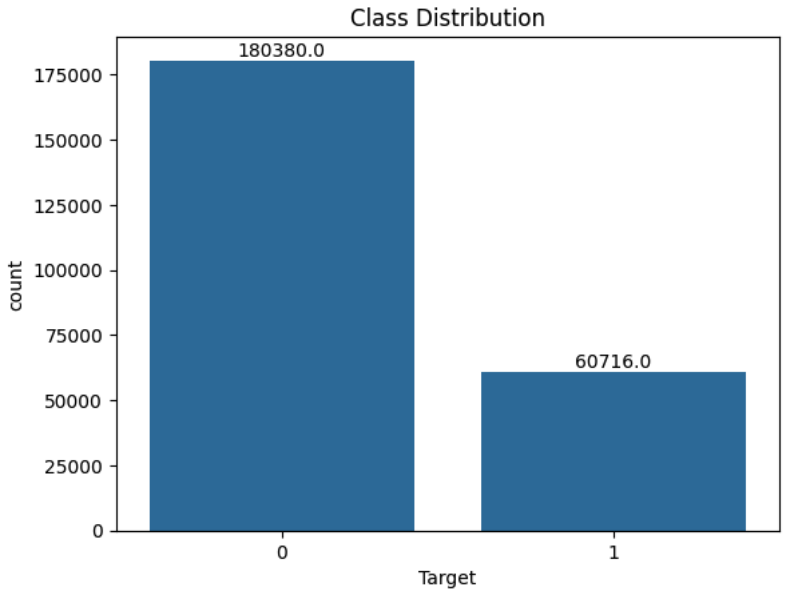


Figure 1: Class Distribution of the Dataset before SMOTE

The minority (delayed flights) class making up only 25.18% of the dataset. This imbalance can affect a model's ability to accurately predict the occurrence of the minority class because it could be biased towards the majority class (not delayed flights).

Balancing the data:

To address the unbalanced issue of the two classes, we used Synthetic Minority Over-Sampling Technique (SMOTE). This technique synthetically resamples the minority class in order to balance the dataset. The distribution of classes after we applied SMOTE is shown in Figure 2. By using SMOTE, we were ensured that the models were no longer biased toward the majority class, which lead to more reliable and fair predictions.

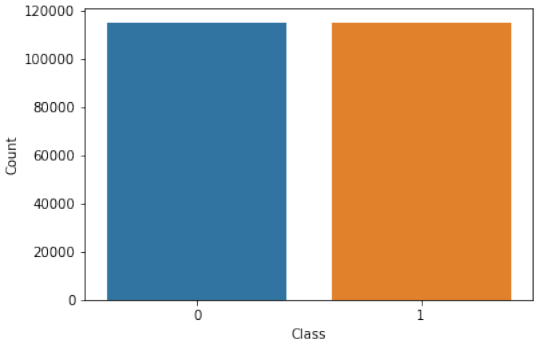


Figure 2: Class Distribution of the Dataset after SMOTE

4 Methodology

One of our research goals is to create a model that will accurately classify whether a flight delay will occur, so the methodology our project will employ is predictive. We will be using supervised statistical modeling approaches that are both parametric, such as logistic regression, and non-parametric, such as Support Vector Machines, Random Forest, Gradient Boosting Machine, Decision Trees, and CatBoosting. Exploring both parametric and non-parametric approaches will allow us to weigh the benefits and drawbacks of the interpretability, flexibility, and performance of a variety of models.

When using machine learning models, splitting your data into training, validation and testing is crucial in order to prevent overfitting. We used the common guideline suggested by data scientists to split our data into three sections. We allocated the data from years 2014-2019 (70 % of the data) to the training set, the data from years 2020-2022 (20 % of the data) to the validation set, and the data from years 2023-2024 to the testing set (10 % of the data).

In addition to splitting the data into training, validation, and testing sets, we further split our data by state and by season. The seasons were represented by four months: January for winter, April for spring, July for summer, and October for fall. In the end, our data was split in 24 ways: by 2 states, by 4 seasons, and by our three folds (weather-only predictor fold, flight-only predictor fold, weather and flight predictors fold). This way, we were able to examine the effect of weather data in several different contexts.
The purpose of our study was to classify whether a flight will be delayed or not. For this case, we leveraged the following classification models using Python:

4.1 Baseline Model: Intercept-Only Logistic Regression (LR)

A LR model performs a binary classification by predicting which class the data has a higher probability of belonging to. We first leveraged an intercept-only LR model that estimated the probability of a flight delay based solely on the overall class distribution, not taking into account any predictors. By doing so, we established an accuracy level to compare the performance of our more complex classification models to.

4.2 Full Logistic Regression (LR)

Additionally, we fit a full LR model with predictors to classify flight delays. This method of parametric modeling assumes a linear relationship between the predictors and the log odds of the outcome, and is therefore more easily interpretable. That said, LR is more prone to bias due to its sensitivity to anomalies and its highly generalizable nature.

4.3 Linear Support Vector (LinearSVC)

With millions of observations in our dataset, we selected Linear Support Vector Classification (LinearSVC) instead of the standard SVC model due to its efficiency in both speed and memory usage, which makes it much more suitable for handling large datasets. LinearSVC works by finding the most effective dividing line, or "hyperplane," that separates delayed flights from non-delayed ones, aiming to maximize the distance between this boundary and the closest data points. This approach helps enhance the model's accuracy and stability. Unlike standard SVC, which uses a more complex optimization process, LinearSVC focuses on a simpler objective function, making it faster to compute on larger datasets. Additionally, LinearSVC provides interpretable results by highlighting which features have the greatest influence on predicting flight delays, offering valuable insights into the factors that contribute to these delays.

4.4 Decision Tree (DT)

Another method used to classify a flight was delayed or not was using a Decision Tree (DT). DT is a supervised machine learning algorithm that is used for both classification and regression. Kim and Park [2] recommends this method due to its versatility and interpretability. The resulting model, has a tree structure, where each node represents a decision based on a particular feature from the data, and each branch represents the outcome of that. The tree starts from the first node also known as the root node, and recursively splits into further nodes based on the data. The recursive process is continued until the model reaches the leaf node which corresponds to the final prediction.

A common issue with DT is overfitting. We used grid search to test and tune different hyperparameters in order to avoid overfitting and optimize its performance.

4.5 Random Forest (RF)

The RF model combines the outputs of multiple DT models and leverages each of their results to perform both classification and regression predictions. This method of machine learning randomly selects a subset of features to build each decision tree; a final prediction will be determined by majority vote of all of the decision trees' predictions. By averaging predictions across multiple decision trees formulated by randomized features, variance is reduced and accuracy is often improved when compared to DT models. This is advantageous when utilizing large datasets with high-dimensional features.

4.6 CatBoost (CatBoost)

CatBoost, a gradient boosting algorithm optimized for handling categorical features, and it was chosen to classify flight delays due to its capability to process high-dimensional data without extensive preprocessing. Unlike traditional boosting methods, CatBoost implements ordered boosting, which reduces prediction bias and mitigates overfitting by training trees sequentially in a way that prevents data leakage. The algorithm builds multiple decision trees that work together, with each tree correcting errors from previous trees, resulting in improved predictive accuracy. Hyperparameters such as the learning rate, tree depth, and L2 regularization were carefully tuned to achieve an optimal balance between model complexity and accuracy. Additionally, class weights were applied to address the imbalance between delayed and non-delayed flights, allowing CatBoost to learn effectively from both classes. By setting early stopping rounds, we monitored performance on the validation set to further prevent overfitting. This makes CatBoost particularly suited for high-cardinality features and imbalanced datasets, offering a powerful, interpretable solution for classification tasks.

5 Results

In this section we discuss the result of our analysis by leveraging machine learning models mentioned in the Methodology Section. The tables showing our results can be found in the Appendix. Each page shows the training, validation, and testing results for our five models, and each page shows the results for a particular state/season combination. Additionally, we have four performance metrics listed: overall accuracy, recall, precision, and F-1 score.

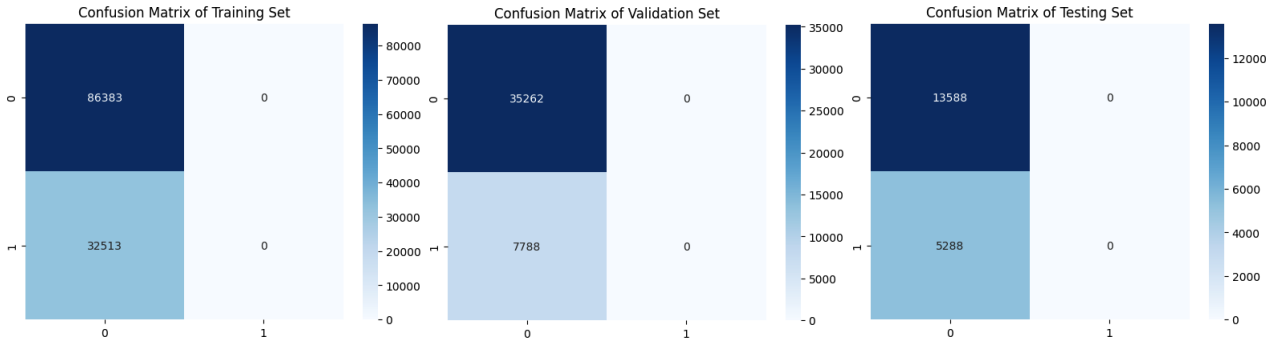
Since our study is focused on classifying delayed flights, we decided to examine the positive predictive value (precision), true positive rate (recall or sensitivity), and F1-scores for each model in addition to overall accuracy. High precision would indicate that when the model predicts a delayed flight it is often correct, and high recall would indicate that the model correctly identifies most delayed flights in our dataset. The F1-score is the harmonic mean of precision and recall, so a high F1-score would indicate that the model has a good balance between precision and recall.

In addition to these measures, we also consider the Receiver Operating Characteristic (ROC) curve and feature importance to evaluate model performance using the testing data. The ROC curve plots the true positive rate (sensitivity) against the true negative rate (specificity) at various thresholds, helping us visualize how well the model distinguishes between delayed and non-delayed flights. A higher area under the ROC curve (AUC) indicates better overall classification performance. Feature importance, on the other hand, identifies which input variables (e.g., weather conditions, flight distance) have the most influence on the model's predictions, helping us understand what factors contribute most to flight delays. More information on the feature importance for our models can be found in the Appendix as well.

Below, we went in-depth on our results for January in Illinois using all of our predictors as an example. Any other results can be found in the Appendix.

5.1 Baseline Model: Intercept-Only Logistic Regression (LR)

First, we fit the balanced training data to an intercept-only LR model in order to establish a baseline measurement for prediction accuracy to compare our subsequent models to. As you can see in Tables 2-4, the baseline model had relatively high overall accuracy but received a 0 in precision, recall, and F1-score. Since the model was trained on a balanced dataset, the classification threshold defaulted to 50%; this meant that when testing on the unbalanced training, validation, and testing datasets the model predicted the majority class every time. Recall that the class distribution for our dataset as a whole is approximately 75% non-delayed flights and 25% delayed flights, so by predicting the majority class every time the overall accuracy reflects the class distribution of the unbalanced sets. By doing this we essentially created a model with no predictive power, which is reflected in the ROC curve in Figure 4; the AUC is 0.5 which indicates that this model essentially predicts flight delays by random chance. Ideally, our more complex classification models will achieve a higher AUC score and perform better than random chance.



(a) Training Set (b) Validation Set (c) Test set

Figure 3: Confusion Matrix for training, validation and testing set

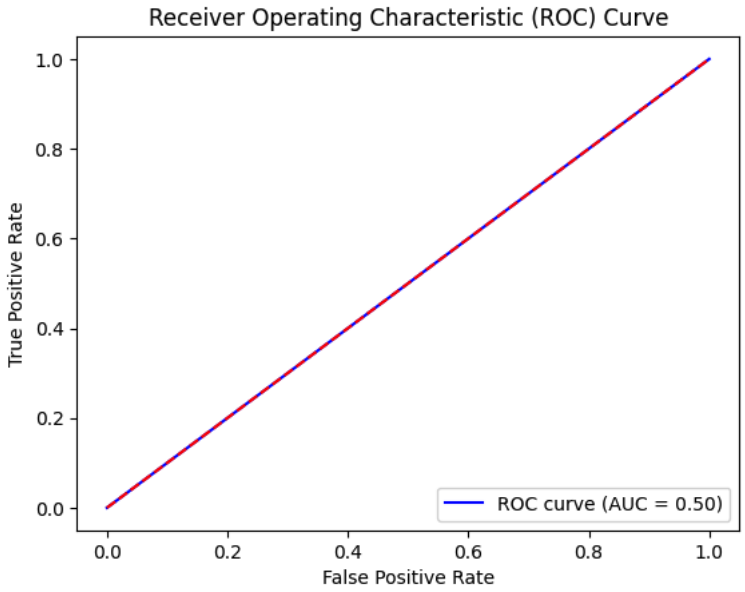
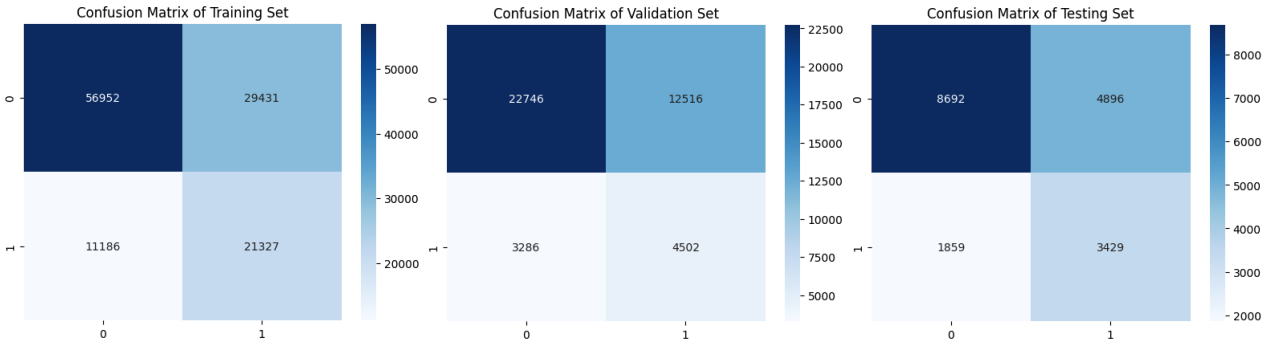


Figure 4: Baseline LR ROC Curve

5.2 Full Logistic Regression (LR)

Next, we fit a full LR model to evaluate the effectiveness of a parametric model that assumes a linear relationship between the weather predictors and the occurrence of flight delays. The precision, recall, and F1-scores for the full LR model seen in Tables 2-4 are significantly higher than that of the baseline model, meaning that this model was able to accurately classify a larger percentage of delayed flights and performed better than random chance predictions.



(a) Training Set (b) Validation Set (c) Test set

Figure 5: Confusion Matrix for training, validation and testing set

Although the overall accuracy decreased, the trade-off of having higher precision and recall indicates better overall classification performance as indicated in the ROC curve in Figure 6. Our full LR model was able to achieve an AUC of 0.71, indicating that there is some predictive power in the model's ability to distinguish between delayed and non-delayed flights, but there is certainly room for improvement.

The feature importance graph in Figure 7 shows the coefficient value of the most significant predictors in this model. These values indicate that weather and temporal predictors had the most influence on the model, demonstrating that weather predictors do have a significant effect on prediction of flight delays.

Given the improvement in classification performance in the full LR model compared to the baseline model, this indicates that there is some linear relationship between the predictors with the highest coefficient values and the log-odds of flight delay occurrence. We now move on to our non-parametric models to determine if a more flexible classification model will perform better than one that assumes linearity.

5.3 Linear Support Vector (LinearSVC)

The preliminary results of our model show moderate success in distinguishing between delayed and non-delayed flights, with a validation accuracy of 80% and a test accuracy of 74%. However, accuracy alone does not fully capture the

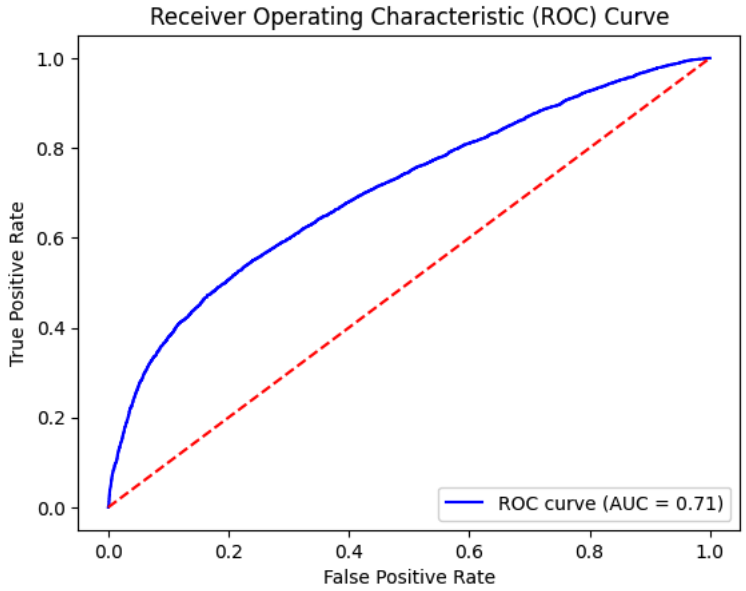


Figure 6: Full LR ROC Curve

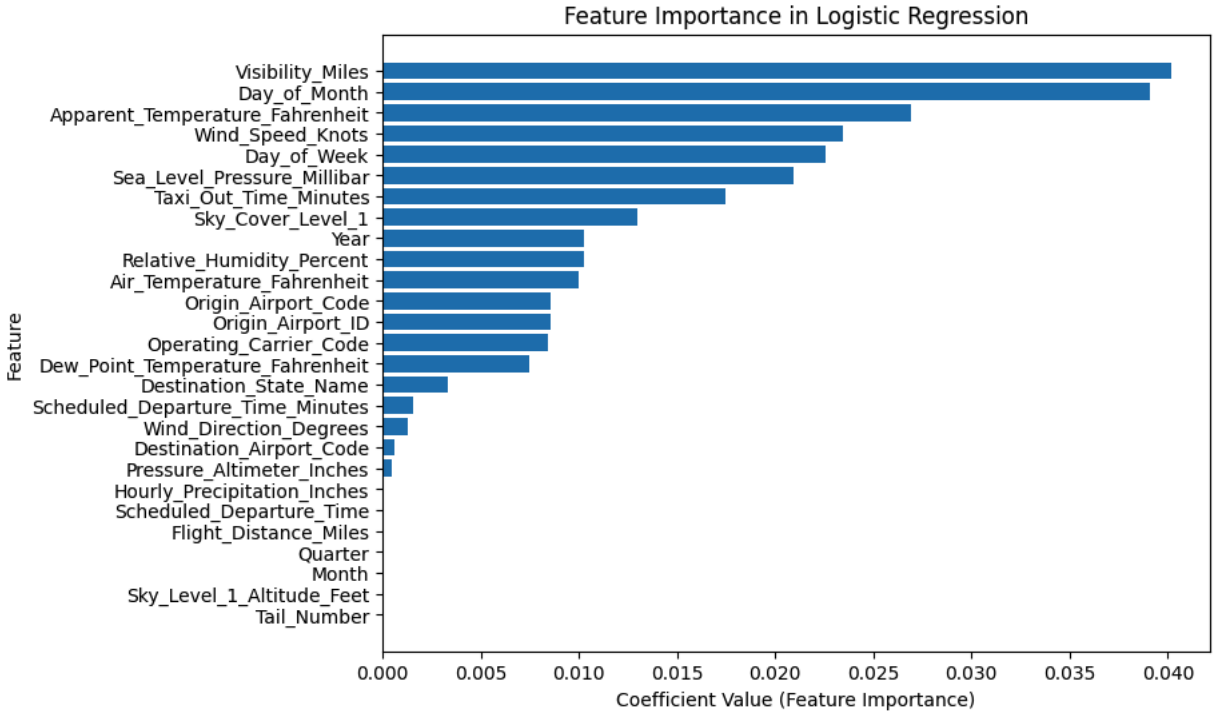
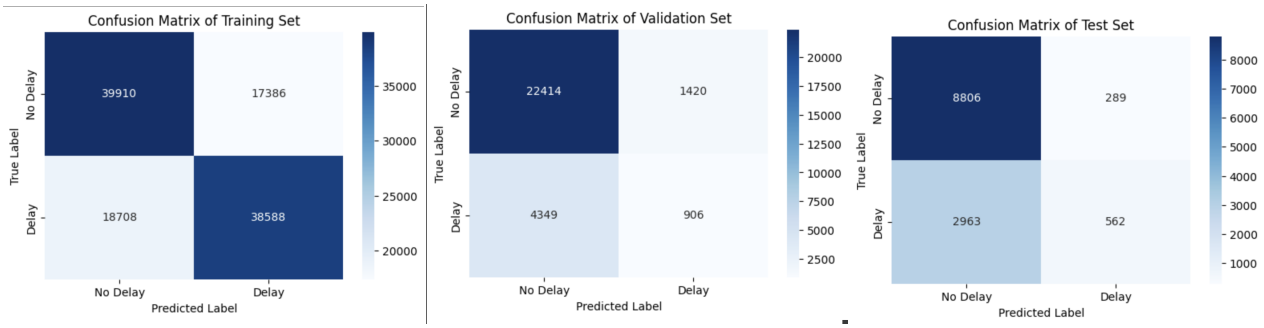


Figure 7: Full LR Feature Importance

model's limitations, especially in identifying delayed flights. While the model performs well for non-delayed flights, achieving high precision and recall for this majority class, it struggles with the minority class of delayed flights. This is evident in the low recall and F1-score for delayed flights, suggesting that the model frequently misses these cases.

The confusion matrices in Figure 8 show that the model is good at predicting flights that won't be delayed across all sets, but it struggles to accurately iden-



(a) Training Set (b) Validation Set (c) Test set

Figure 8: Confusion Matrix for training, validation and testing set

tify delayed flights, especially as it moves from training to validation to test data. This trend suggests that while the model learns well from the training data, it has difficulty generalizing, particularly for predicting delays, indicating it could benefit from adjustments that are listed in a future paragraph to improve accuracy on real-world data.

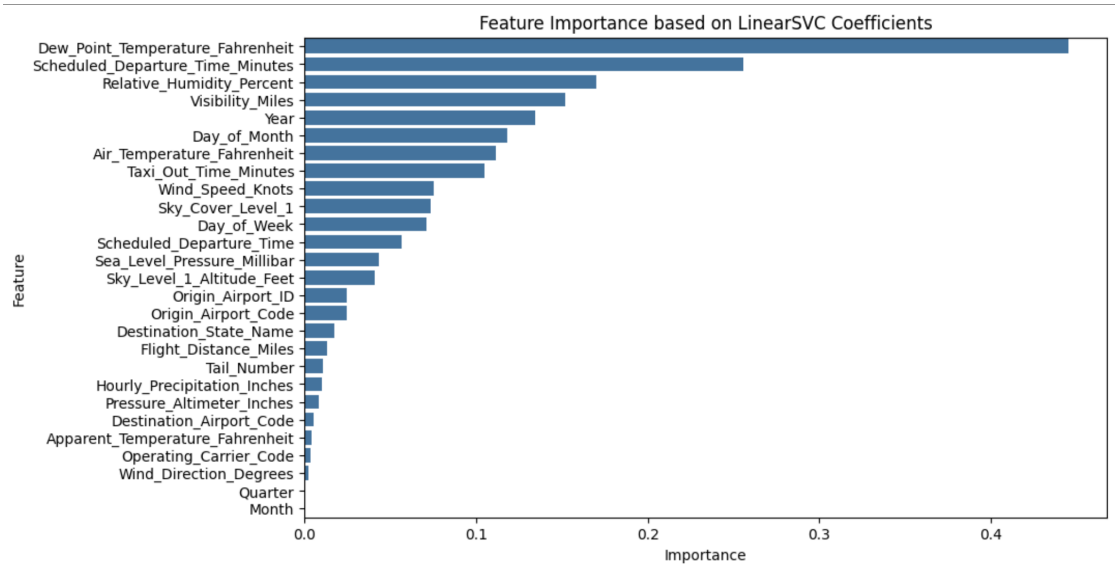


Figure 9: LinearSVC Feature Importance

The feature importance results in Figure 9 show which factors most influence the model in predicting flight delays, based on the training data. Key factors like dew point temperature, scheduled departure time, and humidity stand out, suggesting weather and timing significantly impact delays. We used the training set for this analysis to ensure that the model's learning process is separate from the validation and test sets, helping us avoid data leakage. Data leakage occurs when information from the validation or test sets unintentionally influences the model during training, leading to overly optimistic performance metrics that won't generalize to unseen data. This approach allows us to see which features the model relies on to make predictions without impacting its performance on new, unseen data.

Moving on, lets explore the results of the ROC Curve in Figure 10 (based on the test set):

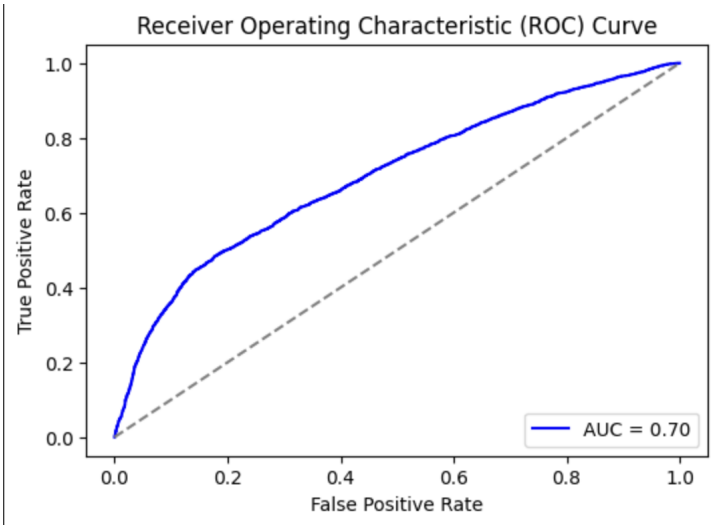


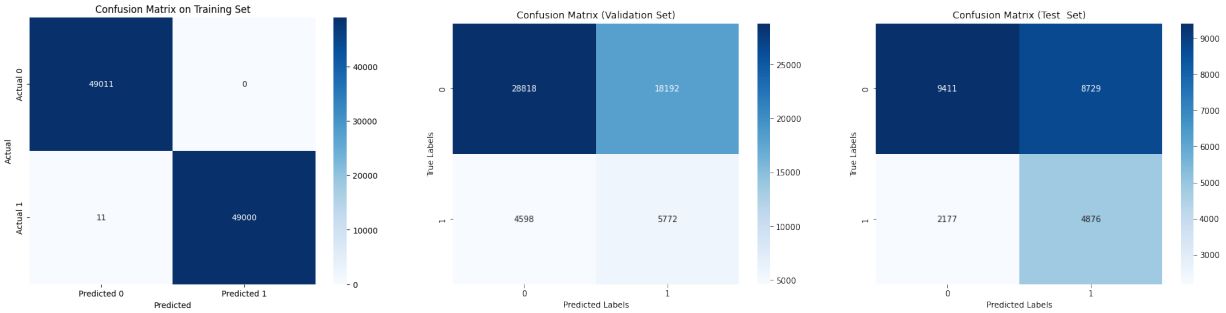
Figure 10: LinearSVC ROC Curve

The ROC curve in Figure 10 shows that the model has moderate discriminatory power with an Area Under the Curve (AUC) of 0.70. This indicates that the model can distinguish between delayed and non-delayed flights better than random guessing (an AUC of 0.5) but still has room for improvement. The curve's gradual rise suggests that the model achieves a decent true positive rate but at the cost of a higher false positive rate, especially as it tries to capture more delayed flights. Overall, while the model shows some ability to predict delays, further optimization is needed to improve its reliability.

Although SMOTE was used to balance the dataset during training, issues with recall for delays persist. This indicates potential for improvement, and additional techniques, such as experimenting with alternative sampling methods (like ADASYN or Borderline-SMOTE), adjusting the decision threshold, or exploring cost-sensitive learning, could enhance the model's performance. These results suggest that the selected methodology has the potential to address the research question of predicting flight delays but may need further refinement to achieve reliable results for real-world application. We are optimistic about improving recall and achieving a better balance across classes through these next steps, which will help ensure the model's predictions are both actionable and accurate.

5.4 Decision Tree (DT)

The decision tree model appears to be overfitting the training data, achieving perfect scores on the training dataset. This may be due to the dataset imbalance, which was addressed using SMOTE.



(a) Training Set (b) Validation Set (c) Testing Set

Figure 11: Confusion Matrix for training, validation and testing set

The overfitting is most obvious in Figure 11(a). There are 0 false positives, while the count of false negatives is 11; hence, this model is almost perfect on a training set. However, when fitting the model to the validation set that contained unbalanced data, the false positives went as high as 18192 and the number of false negatives as high as 4598. Such an increase already hints at poor generalization of the model beyond the training data.

This translates into 8729 false positives and 2177 false negatives when applied to unseen, real-world data or the testing set; that is, it does better but still can make mistakes. Fitting to unseen data decreases the accuracy of the model; but this is to be expected, however this poor performance on testing shows that the model is now less over fitted and can generalize better on the real-world data.

The reason that the model was less over fitted in the validation and testing set was because we used Grid Search [10] to find the best parameters. Table 5, shows the different parameter used to find the best and optimal tree.

Hyperparameter	**Values**
`max_leaf_nodes`	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 50, 100
`min_impurity_decrease`	0.0001, 0.0005, 0.001, 0.005, 0.01
`max_depth`	1, 3, 5
`min_samples_split`	2, 3, 4, 5, 6

Table 2: Hyperparameter Grid Search Values

After using the Grid Search to find the optimal based on the hyperparameter values from Table 5, we found the following values to result in the most optimal tree shown in Table 6. Lastly, we were also interested in which variables were the most important for classifying flight delays based on the decision tree. We used the feature importance from the decision tree to plot the most important features based on the decision tree model as shown in Figure 12.

Hyperparameter	Optimal Value
max_leaf_nodes	6
min_impurity_decrease	0.001
max_depth	5
min_samples_split	4

Table 3: Optimal Hyperparameter Values for Decision Tree

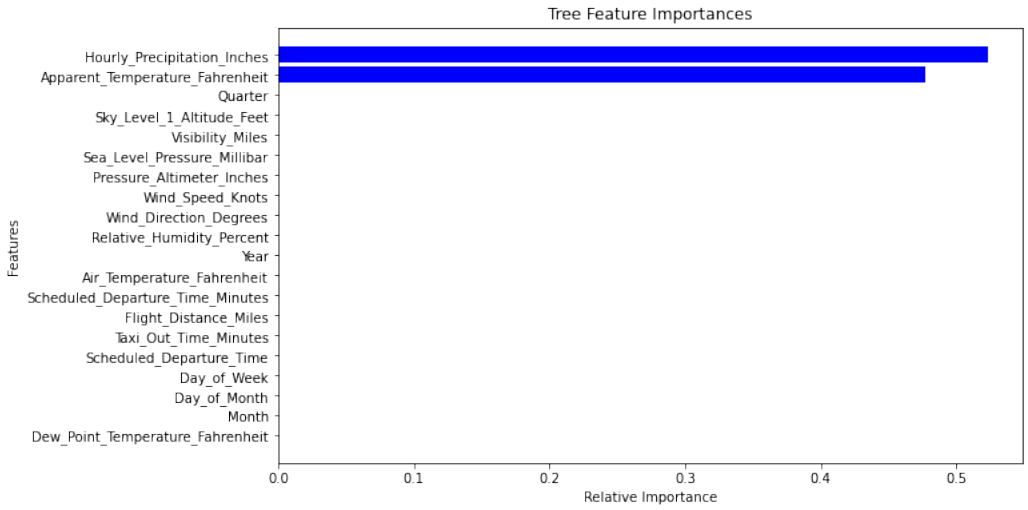


Figure 12: Tree Feature Importance

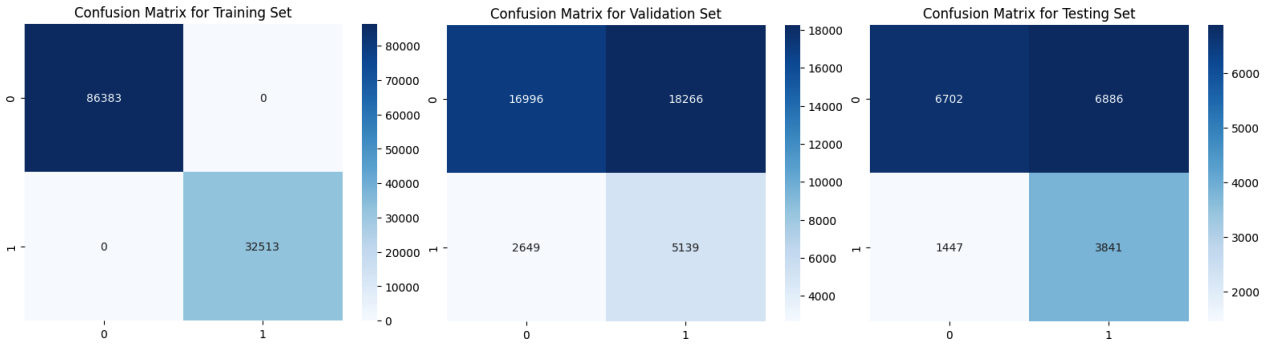
Based on Figure 12, the most important features according to the decision tree model are hourly precipitation in inches and apparent temperature in Fahrenheit.

5.5 Random Forest (RF)

Similar to the DT model, the RF model was overfit to the balanced training data which resulted in perfect scores when testing on the unbalanced training set. For the validation and testing sets however, the most notable difference in our results was that the precision performed the worse of all of the models, and the recall performed the best. This meant that when the model would classify a flight as delayed it was often not correct, but it was able to correctly classify the highest percentage of delayed flights in the dataset compared to other models.

The feature importance graph is Figure 14 shows that the predictors shared more similar coefficient values compared to other models. This suggests that though there are more significant features in the model, they did not stand out in influencing the model's predictive power. Something else to note was that the model ranked non-weather predictors such as Scheduled Departure Time and Tail Number as highly important, while weather predictors were often found to be more significant in other models.

As a result of the imbalance between precision and recall, the AUC was lower than most models at a rate of 0.65. As you can see in the ROC curve in



(a) Training Set (b) Validation Set (c) Testing Set

Figure 13: Confusion Matrix for training, validation and testing set

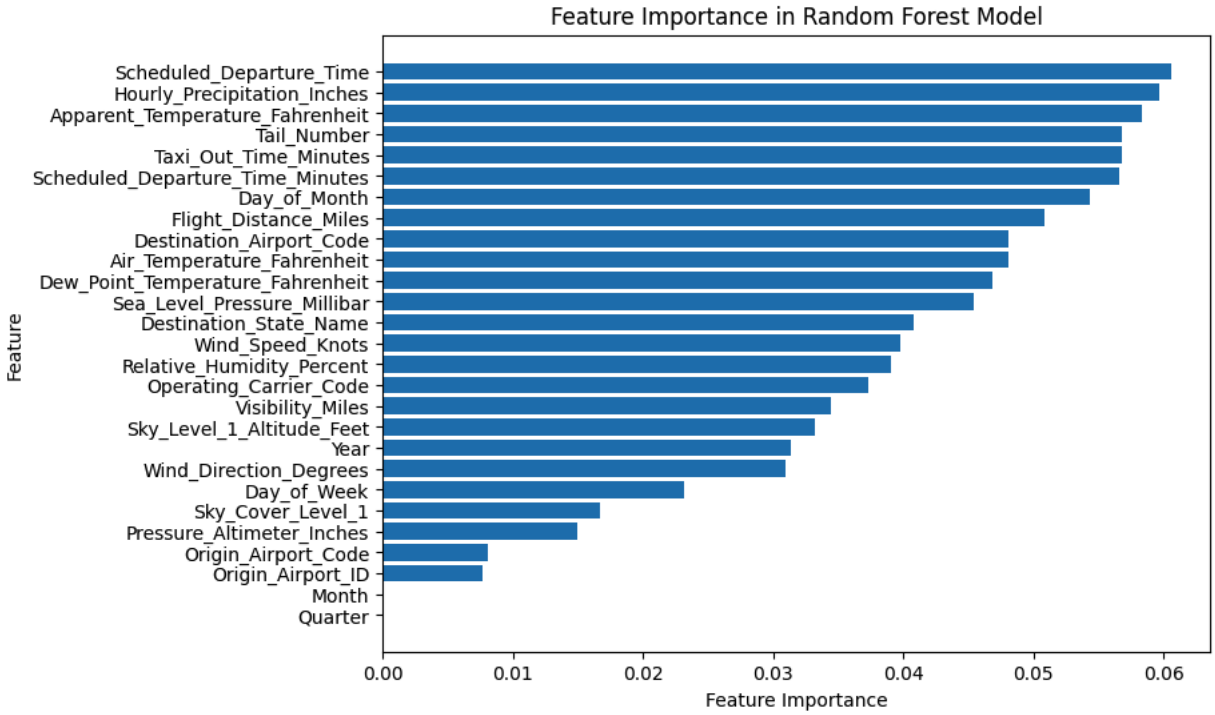


Figure 14: RF Feature Importance

Figure 15, there was little advantage in the trade-off between sensitivity and specificity and as a result the model only performed marginally better in overall classification performance than the baseline model.

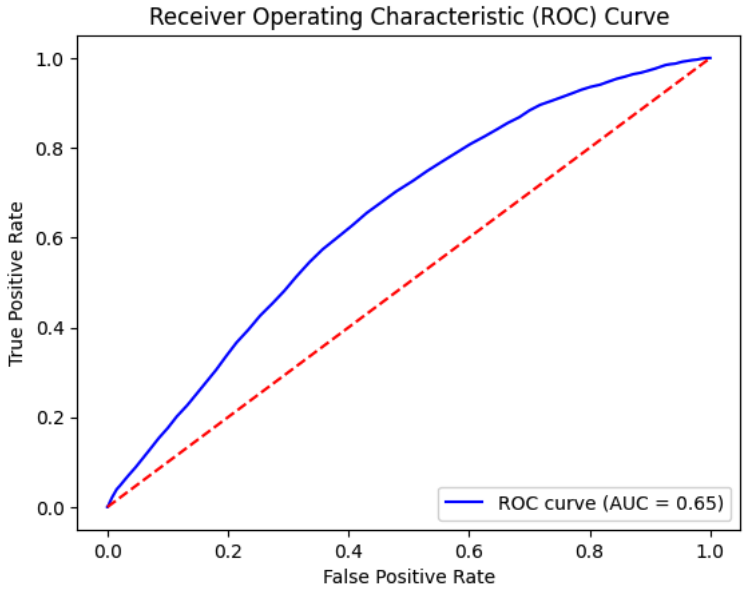


Figure 15: RF ROC Curve

5.6 CatBoost (CatBoost)

CatBoost, a gradient boosting algorithm optimized for handling categorical variables efficiently, was chosen for predicting flight delays due to its ability to process high-dimensional data without extensive preprocessing. Unlike traditional boosting methods, CatBoost incorporates ordered boosting, which mitigates overfitting and reduces prediction bias. Key hyperparameters such as the learning rate (0.05), tree depth (8), and L2 regularization (2) were tuned to balance model complexity and accuracy. To address the class imbalance between delayed and non-delayed flights, we applied class weights, enhancing the model's ability to learn from the minority class. Early stopping rounds were set to 200, ensuring the model did not overfit on the validation set. CatBoost achieved strong performance metrics, with an F1-score of 0.37 on the validation set, and F1-score of 0.47 on the testing set showing its capability to capture complex patterns in flight delay prediction. Its interpretability and efficiency in processing categorical data make CatBoost a robust choice for classification in this high-dimensional context.

Now lets look at feature importance to see which variables were most influential to our model:

The ROC curve for CatBoost model can be seen in figure 17:

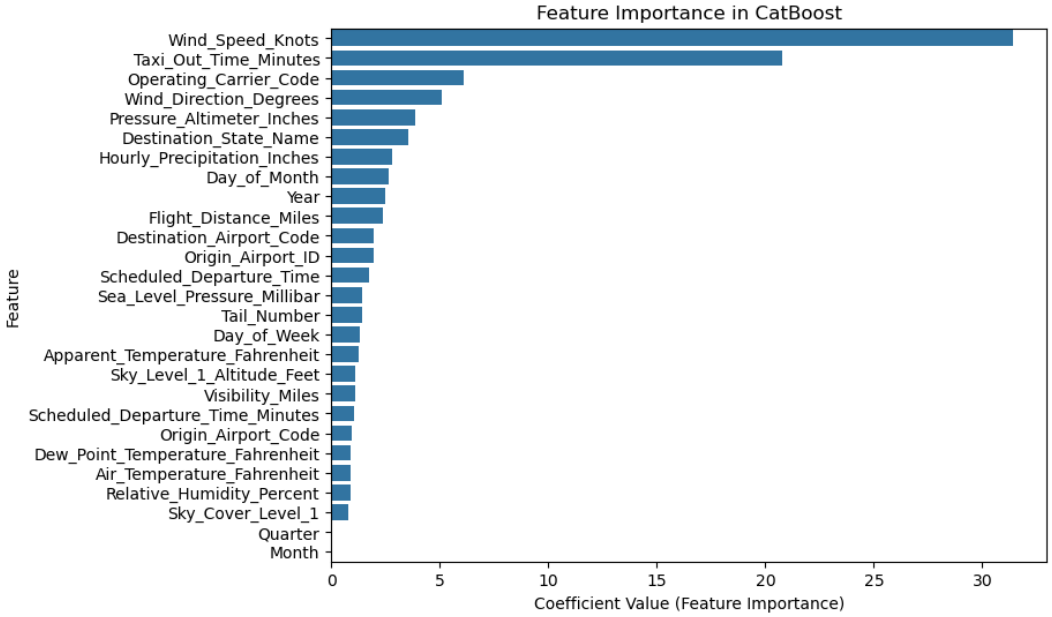


Figure 16: CatBoost Feature Importance

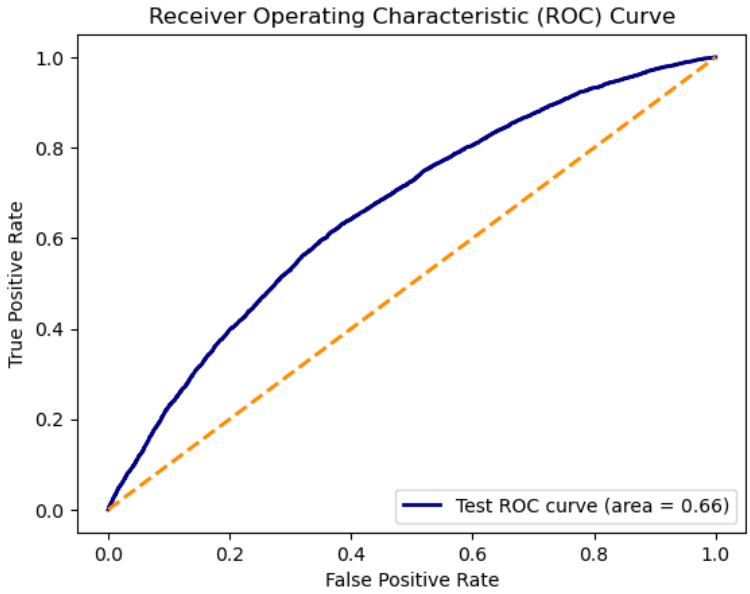


Figure 17: CatBoost ROC Curve

ROC curve demonstrate the model's performance in distinguishing between classes. The True Positive Rate (TPR) is plotted against the False Positive Rate (FPR) for various classification thresholds. The Area Under the Curve (AUC) of 0.66 indicates moderate discriminatory ability, with the blue curve representing the model's performance and the orange diagonal line serving as a baseline for random guessing (AUC = 0.5).

6 Conclusion

This study highlights the importance of incorporating weather data into flight delay prediction models and demonstrates how combining it with flight-specific data produces more reliable results. Weather data alone proved highly effective at identifying delayed flights, particularly during seasons like spring when delays are more frequent. The high recall achieved in these cases reflects the model's ability to capture the majority of delayed flights, which is critical for minimizing the negative effects of missed predictions—such as costly re-scheduling, passenger dissatisfaction, and environmental impacts from increased fuel consumption.

To evaluate the model's performance, we focused on Recall and F1 score because they address the unique challenges of flight delay prediction. Recall measures how well the model identifies delayed flights, which is crucial in reducing the consequences of missed delays (false negatives). However, focusing solely on Recall can increase false positives—flights flagged as delayed when they're actually on time. This is where the F1 score becomes critical, as it balances Recall with Precision to ensure the model is both accurate and efficient. By prioritizing these metrics, we ensure the model aligns with real-world needs, helping airlines identify delays without creating unnecessary disruptions from false alarms.

While weather data alone performs well in capturing delays, it does not account for flight-specific details like schedules, air traffic, or operational factors, leading to a higher number of false positives. By integrating flight and weather data, the model captures both broad trends influenced by weather and specific conditions that impact delays. This combined approach improves the F1 score, resulting in a model that reliably predicts both delayed and on-time flights. Such a balance is crucial for practical applications, ensuring stakeholders can trust the model's predictions.

In conclusion, this study demonstrates that using only weather data provides high Recall, meaning the model can identify most delayed flights, which is particularly useful in anticipating widespread disruptions during adverse weather conditions. However, combining flight-specific and weather data results in a higher F1 score, which balances the need to catch delayed flights with minimizing false alarms. In real-world terms, this means that while weather data alone helps airlines prepare for potential delays, the combined model offers more accurate and actionable predictions. By reducing unnecessary disruptions and focusing resources on the most critical delays, airlines can improve operational efficiency, passenger satisfaction, and environmental outcomes. This integration of diverse data sources represents a powerful approach to managing delays more effectively.

7 Future Work

Moving forward, further research on how effective weather predictors are at predicting the duration of flight delays could prove useful. Based on our literature review, current research focuses mainly on classifying flight delays rather than estimating their duration. Using regression models to predict the number of minutes a flight will be delayed would be a valuable resource to the airline industry as it could help determine the cascading effect the flight delay will have on subsequent flights. Furthermore, due to time constraints

our study was only able to focus on weather and flight data from two states. Future researchers could potentially come to interesting conclusions by integrating the weather patterns of the different regions of the United States.

8 Appendix

8.1 Results for Accuracy Measures

1. Only Weather Predictors

   a) Illinois

      i) January

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.65	0.40	0.61	0.49
Linear SVC	0.63	0.64	0.60	0.62
DT	0.62	0.59	0.73	0.65
RF	1	1	1	1
CatBoost	0.61	0.57	0.91	0.70

Table 4: Performance metrics on the training set using weather predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.64	0.25	0.50	0.34
Linear SVC	0.65	0.26	0.51	0.35
DT	0.66	0.18	0.39	0.24
RF	0.57	0.21	0.52	0.30
CatBoost	0.37	0.19	0.82	0.32

Table 5: Performance metrics on the validation set using weather predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.64	0.41	0.65	0.51
Linear SVC	0.65	0.42	0.64	0.51
DT	0.49	0.30	0.74	0.43
RF	0.57	0.34	0.60	0.44
CatBoost	0.41	0.30	0.87	0.45

Table 6: Performance metrics on the testing set using weather predictors

ii) April

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.60	0.26	0.55	0.36
Linear SVC	0.59	0.60	0.56	0.58
DT	0.57	0.66	0.30	0.41
RF	1	1	1	1
CatBoost	0.52	0.51	0.97	0.67

Table 7: Performance metrics on the training set using weather predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.54	0.07	0.44	0.12
Linear SVC	0.54	0.06	0.42	0.11
DT	0.73	0.07	0.22	0.10
RF	0.45	0.07	0.51	0.12
CatBoost	0.10	0.07	0.96	0.13

Table 8: Performance metrics on the validation set using weather predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.35	0.25	0.85	0.38
Linear SVC	0.28	0.23	0.90	0.37
DT	0.25	0.24	0.99	0.38
RF	0.33	0.24	0.86	0.38
CatBoost	0.23	0.23	0.99	0.38

Table 9: Performance metrics on the testing set using weather predictors

iii) July

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.58	0.34	0.58	0.43
Linear SVC	0.59	0.60	0.60	0.60
DT	0.50	1	0	0
RF	1	1	1	1
CatBoost	0.57	0.54	0.92	0.68

Table 10: Performance metrics on the training set using weather predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.47	0.16	0.63	0.26
Linear SVC	0.56	0.19	0.60	0.29
DT	0.85	1.0	0	0
RF	0.47	0.16	0.60	0.25
CatBoost	0.26	0.15	0.90	0.27

Table 11: Performance metrics on the validation set using weather predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.64	0.47	0.55	0.51
Linear SVC	0.62	0.46	0.65	0.54
DT	0.66	1.0	0	0
RF	0.59	0.43	0.62	0.51
CatBoost	0.49	0.39	0.88	0.54

Table 12: Performance metrics on the testing set using weather predictors

iv) October

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.59	0.24	0.56	0.34
Linear SVC	0.58	0.59	0.55	0.57
DT	0.61	0.65	0.48	0.55
RF	1	1	1	1
CatBoost	0.53	0.52	0.97	0.68

Table 13: Performance metrics on the training set using weather predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.63	0.19	0.42	0.26
Linear SVC	0.64	0.20	0.45	0.27
DT	0.75	0.26	0.32	0.29
RF	0.62	0.16	0.35	0.22
CatBoost	0.23	0.16	0.93	0.27

Table 14: Performance metrics on the validation set using weather predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.71	0.18	0.24	0.21
Linear SVC	0.72	0.20	0.29	0.24
DT	0.77	0.17	0.11	0.14
RF	0.70	0.17	0.23	0.19
CatBoost	0.16	0.15	0.98	0.27

Table 15: Performance metrics on the testing set using weather predictors

b) Georgia

i) January

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.62	0.26	0.55	0.36
Linear SVC	0.60	0.61	0.55	0.58
DT	0.62	0.65	0.52	0.58
RF	1	1	1	1
CatBoost	0.53	0.52	0.96	0.68

Table 16: Performance metrics on the training set using weather predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.71	0.18	0.36	0.24
Linear SVC	0.70	0.17	0.36	0.23
DT	0.73	0.16	0.40	0.23
RF	0.52	0.14	0.57	0.23
CatBoost	0.25	0.13	0.90	0.23

Table 17: Performance metrics on the validation set using weather predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.58	0.25	0.50	0.33
Linear SVC	0.59	0.25	0.51	0.33
DT	0.69	0.25	0.35	0.29
RF	0.47	0.23	0.64	0.33
CatBoost	0.31	0.22	0.91	0.36

Table 18: Performance metrics on the testing set using weather predictors

ii) April

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.58	0.23	0.58	0.33
Linear SVC	0.60	0.61	0.57	0.59
DT	0.59	0.68	0.33	0.45
RF	1	1	1	1
CatBoost	0.55	0.52	0.95	0.68

Table 19: Performance metrics on the training set using weather predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.60	0.10	0.46	0.16
Linear SVC	0.57	0.10	0.55	0.18
DT	0.75	0.11	0.30	0.17
RF	0.41	0.09	0.68	0.16
CatBoost	0.19	0.08	0.89	0.15

Table 20: Performance metrics on the validation set using weather predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.57	0.29	0.54	0.38
Linear SVC	0.60	0.29	0.47	0.36
DT	0.69	0.28	0.19	0.23
RF	0.49	0.26	0.61	0.37
CatBoost	0.31	0.24	0.92	0.39

Table 21: Performance metrics on the testing set using weather predictors

iii) July

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.54	0.28	0.70	0.40
Linear SVC	0.60	0.59	0.68	0.63
DT	0.57	0.54	0.85	0.67
RF	1	1	1	1
CatBoost	0.62	0.60	0.92	0.71

Table 22: Performance metrics on the training set using weather predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.54	0.18	0.66	0.28
Linear SVC	0.57	0.18	0.63	0.28
DT	0.38	0.16	0.82	0.26
RF	0.46	0.16	0.71	0.26
CatBoost	0.37	0.16	0.84	0.27

Table 23: Performance metrics on the validation set using weather predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.58	0.39	0.68	0.50
Linear SVC	0.61	0.41	0.63	0.50
DT	0.53	0.36	0.72	0.48
RF	0.51	0.35	0.71	0.47
CatBoost	0.49	0.36	0.83	0.50

Table 24: Performance metrics on the testing set using weather predictors

iv) October

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.57	0.17	0.60	0.27
Linear SVC	0.59	0.59	0.61	0.60
DT	0.58	0.58	0.61	0.59
RF	1	1	1	1
CatBoost	0.54	0.52	0.95	0.68

Table 25: Performance metrics on the training set using weather predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.55	0.11	0.55	0.19
Linear SVC	0.55	0.12	0.58	0.20
DT	0.44	0.11	0.69	0.19
RF	0.49	0.10	0.56	0.17
CatBoost	0.18	0.10	0.94	0.18

Table 26: Performance metrics on the validation set using weather predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.55	0.16	0.52	0.25
Linear SVC	0.55	0.17	0.55	0.26
DT	0.63	0.17	0.39	0.23
RF	0.44	0.15	0.61	0.24
CatBoost	0.20	0.14	0.93	0.25

Table 27: Performance metrics on the testing set using weather predictors

2. Only Flight Predictors

a) Illinois

i) January

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.60	0.36	0.58	0.45
Linear SVC	0.63	0.63	0.64	0.64
DT	0.59	0.75	0.27	0.40
RF	1	1	1	1
CatBoost	0.61	0.59	0.92	0.70

Table 28: Performance metrics on the training set using flight predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.53	0.20	0.52	0.29
Linear SVC	0.77	0.19	0.09	0.12
DT	0.86	0	0	0
RF	0.63	0.20	0.33	0.25
CatBoost	0.33	0.18	0.77	0.30

Table 29: Performance metrics on the validation set using flight predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.56	0.31	0.46	0.37
Linear SVC	0.71	0.15	0.00	0.01
DT	0.74	0.0	0.0	0.0
RF	0.64	0.34	0.30	0.32
CatBoost	0.46	0.32	0.81	0.46

Table 30: Performance metrics on the testing set using flight predictors

ii) April

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.58	0.26	0.55	0.35
Linear SVC	0.60	0.60	0.61	0.61
DT	0.58	0.55	0.79	0.65
RF	1	1	1	1
CatBoost	0.75	0.68	0.91	0.78

Table 31: Performance metrics on the training set using flight predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.56	0.08	0.49	0.13
Linear SVC	0.84	0.07	0.12	0.09
DT	0.35	0.07	0.72	0.13
RF	0.76	0.09	0.26	0.13
CatBoost	0.44	0.08	0.67	0.14

Table 32: Performance metrics on the validation set using flight predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.62	0.32	0.56	0.41
Linear SVC	0.76	0.43	0.05	0.08
DT	0.50	0.29	0.78	0.42
RF	0.69	0.31	0.26	0.28
CatBoost	0.57	0.29	0.55	0.38

Table 33: Performance metrics on the testing set using flight predictors

iii) July

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.61	0.37	0.60	0.46
Linear SVC	0.62	0.62	0.63	0.62
DT	0.63	0.61	0.67	0.64
RF	1	1	1	1
CatBoost	0.71	0.65	0.90	0.76

Table 34: Performance metrics on the training set using flight predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.60	0.18	0.48	0.26
Linear SVC	0.69	0.19	0.33	0.24
DT	0.54	0.17	0.53	0.25
RF	0.65	0.15	0.31	0.21
CatBoost	0.40	0.16	0.75	0.27

Table 35: Performance metrics on the validation set using flight predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.63	0.46	0.60	0.52
Linear SVC	0.66	0.51	0.31	0.39
DT	0.60	0.44	0.59	0.50
RF	0.63	0.46	0.41	0.44
CatBoost	0.51	0.39	0.79	0.53

Table 36: Performance metrics on the testing set using flight predictors

iv) October

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.58	0.24	0.56	0.33
Linear SVC	0.60	0.61	0.62	0.61
DT	0.64	0.79	0.37	0.50
RF	1	1	1	1
CatBoost	0.78	0.72	0.91	0.81

Table 37: Performance metrics on the training set using flight predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.57	0.19	0.54	0.28
Linear SVC	0.84	0.13	0.01	0.01
DT	0.85	1.0	0	0
RF	0.71	0.18	0.24	0.20
CatBoost	0.51	0.19	0.70	0.31

Table 38: Performance metrics on the validation set using flight predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.48	0.19	0.71	0.30
Linear SVC	0.84	0.00	0.00	0.00
DT	0.84	1.0	0	0
RF	0.74	0.19	0.20	0.19
CatBoost	0.66	0.19	0.37	0.26

Table 39: Performance metrics on the testing set using flight predictors

b) Georgia

i) January

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.58	0.24	0.57	0.34
Linear SVC	0.61	0.61	0.62	0.61
DT	0.66	0.62	0.83	0.71
RF	1	1	1	1
CatBoost	0.60	0.61	0.92	0.74

Table 40: Performance metrics on the training set using flight predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.54	0.16	0.65	0.26
Linear SVC	0.86	0.16	0.03	0.04
DT	0.90	1.0	0	0
RF	0.74	0.16	0.25	0.20
CatBoost	0.43	0.14	0.70	0.23

Table 41: Performance metrics on the validation set using flight predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.58	0.24	0.49	0.32
Linear SVC	0.79	0.00	0.00	0.00
DT	0.82	1.0	0	0
RF	0.71	0.27	0.23	0.25
CatBoost	0.50	0.25	0.70	0.37

Table 42: Performance metrics on the testing set using flight predictors

ii) April

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.58	0.23	0.56	0.32
Linear SVC	0.59	0.59	0.60	0.59
DT	0.60	0.58	0.72	0.64
RF	1	1	1	1
CatBoost	0.74	0.69	0.90	0.78

Table 43: Performance metrics on the training set using flight predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.54	0.09	0.52	0.16
Linear SVC	0.77	0.09	0.19	0.12
DT				
RF	0.47	0.08	0.53	0.14
CatBoost	0.41	0.09	0.71	0.17

Table 44: Performance metrics on the validation set using flight predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.59	0.31	0.57	0.40
Linear SVC	0.73	0.35	0.12	0.18
DT	0.54	0.28	0.58	0.38
RF	0.66	0.31	0.31	0.31
CatBoost	0.52	0.28	0.63	0.39

Table 45: Performance metrics on the testing set using flight predictors

iii) July

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.62	0.31	0.63	0.42
Linear SVC	0.63	0.63	0.65	0.64
DT	0.67	0.67	0.67	0.67
RF	1	1	1	1
CatBoost	0.66	0.61	0.91	0.73

Table 46: Performance metrics on the training set using flight predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.64	0.21	0.60	0.31
Linear SVC	0.70	0.23	0.50	0.31
DT	0.64	0.20	0.56	0.29
RF	0.70	0.20	0.38	0.26
CatBoost	0.45	0.16	0.76	0.27

Table 47: Performance metrics on the validation set using flight predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.62	0.41	0.61	0.49
Linear SVC	0.66	0.44	0.43	0.44
DT	0.62	0.40	0.47	0.43
RF	0.65	0.43	0.47	0.45
CatBoost	0.52	0.36	0.76	0.5

Table 48: Performance metrics on the testing set using flight predictors

iv) October

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.58	0.16	0.53	0.25
Linear SVC	0.59	0.59	0.59	0.59
DT	0.65	0.60	0.89	0.72
RF	1	1	1	1
CatBoost	0.82	0.78	0.91	0.84

Table 49: Performance metrics on the training set using flight predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.57	0.12	0.55	0.20
Linear SVC	0.80	0.12	0.17	0.14
DT	0.14	0.08	0.48	0.14
RF	0.77	0.12	0.22	0.16
CatBoost	0.55	0.12	0.62	0.21

Table 50: Performance metrics on the validation set using flight predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.48	0.16	0.63	0.26
Linear SVC	0.79	0.15	0.10	0.12
DT	0.49	0.14	0.51	0.22
RF	0.79	0.22	0.20	0.21
CatBoost	0.71	0.21	0.37	0,27

Table 51: Performance metrics on the testing set using flight predictors

3. Flight and Weather Predictors

a) Illinois

i) January

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.66	0.42	0.66	0.51
Linear SVC	0.67	0.68	0.66	0.67
DT	0.65	0.62	0.79	0.69
RF	1	1	1	1
CatBoost	0.76	0.54	0.71	0.61

Table 52: Performance metrics on the training set using all predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.63	0.26	0.58	0.36
Linear SVC	0.80	0.38	0.17	0.23
DT	0.67	0.18	0.39	0.25
RF	0.51	0.22	0.66	0.33
CatBoost	0.69	0.30	0.50	0.37

Table 53: Performance metrics on the validation set using all predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.64	0.41	0.65	0.50
Linear SVC	0.75	0.71	0.16	0.26
DT	0.49	0.30	0.74	0.43
RF	0.56	0.36	0.73	0.48
CatBoost	0.66	0.42	0.53	0.47

Table 54: Performance metrics on the testing set using all predictors

ii) April

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.63	0.29	0.59	0.39
Linear SVC	0.63	0.64	0.61	0.62
DT	0.61	0.70	0.37	0.49
RF	1	1	1	1
CatBoost	0.63	0.58	0.89	0.71

Table 55: Performance metrics on the training set using all predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.47	0.07	0.55	0.13
Linear SVC	0.81	0.06	0.12	0.08
DT	0.73	0.07	0.22	0.10
RF	0.61	0.07	0.35	0.11
CatBoost	0.62	0.08	0.40	0.13

Table 56: Performance metrics on the validation set using all predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.43	0.27	0.87	0.41
Linear SVC	0.73	0.37	0.20	0.26
DT	0.25	0.24	0.99	0.38
RF	0.36	0.25	0.88	0.39
CatBoost	0.36	0.25	0.90	0.40

Table 57: Performance metrics on the testing set using all predictors

iii) July

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.62	0.38	0.60	0.47
Linear SVC	0.63	0.64	0.63	0.63
DT	0.63	0.61	0.73	0.67
RF	1	1	1	1
CatBoost	0.70	0.61	0.92	0.74

Table 58: Performance metrics on the training set using all predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.53	0.18	0.63	0.28
Linear SVC	0.66	0.18	0.39	0.25
DT	0.43	0.16	0.69	0.26
RF	0.53	0.17	0.56	0.26
CatBoost	0.46	0.17	0.73	0.28

Table 59: Performance metrics on the validation set using all predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.61	0.46	0.72	0.56
Linear SVC	0.68	0.55	0.39	0.45
DT	0.58	0.43	0.71	0.54
RF	0.61	0.45	0.62	0.52
CatBoost	0.59	0.44	0.75	0.56

Table 60: Performance metrics on the testing set using all predictors

iv) October

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.62	0.26	0.57	0.63
Linear SVC	0.63	0.65	0.59	0.62
DT	0.67	0.76	0.49	0.59
RF	1	1	1	1
CatBoost	0.68	0.63	0.88	0.74

Table 61: Performance metrics on the training set using all predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.59	0.20	0.56	0.29
Linear SVC	0.81	0.25	0.11	0.15
DT	0.78	0.28	0.28	0.28
RF	0.73	0.23	0.32	0.27
CatBoost	0.67	0.22	0.46	0.30

Table 62: Performance metrics on the validation set using all predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.61	0.20	0.51	0.29
Linear SVC	0.85	0.46	0.03	0.06
DT	0.79	0.18	0.10	0.13
RF	0.78	0.19	0.14	0.16
CatBoost	0.66	0.19	0.36	0.25

Table 63: Performance metrics on the testing set using all predictors

b) Georgia

i) January

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.64	0.28	0.61	0.39
Linear SVC	0.65	0.66	0.63	0.64
DT	0.62	0.65	0.52	0.58
RF	1	1	1	1
CatBoost	0.84	0.81	0.89	0.85

Table 64: Performance metrics on the training set using all predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.63	0.19	0.58	0.28
Linear SVC	0.84	0.24	0.13	0.17
DT	0	0	0	0
RF	0.78	0.17	0.19	0.18
CatBoost	0.84	0.22	0.11	0.15

Table 65: Performance metrics on the validation set using all predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.54	0.24	0.59	0.34
Linear SVC	0.80	0.65	0.06	0.11
DT	0.69	0.25	0.35	0.29
RF	0.73	0.33	0.33	0.33
CatBoost	0.78	0.43	0.16	0.24

Table 66: Performance metrics on the testing set using all predictors

ii) April

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.63	0.26	0.58	0.36
Linear SVC	0.63	0.64	0.60	0.62
DT	0.59	0.70	0.31	0.43
RF	1	1	1	1
CatBoost	0.86	0.85	0.87	0.87

Table 67: Performance metrics on the training set using all predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.53	0.10	0.60	0.18
Linear SVC	0.76	0.11	0.28	0.16
DT	0.83	0.11	0.16	0.13
RF	0.66	0.11	0.45	0.18
CatBoost	0.75	0.12	0.33	0.18

Table 68: Performance metrics on the validation set using all predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.56	0.31	0.67	0.42
Linear SVC	0.73	0.39	0.20	0.27
DT	0.68	0.27	0.21	0.24
RF	0.65	0.32	0.41	0.36
CatBoost	0.71	0.36	0.26	0.31

Table 69: Performance metrics on the testing set using all predictors

iii) July

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.64	0.33	0.64	0.44
Linear SVC	0.65	0.65	0.66	0.65
DT	0.63	0.61	0.75	0.67
RF	1	1	1	1
CatBoost	0.65	0.60	0.90	0.72

Table 70: Performance metrics on the training set using all predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.66	0.23	0.62	0.33
Linear SVC	0.72	0.25	0.55	0.34
DT	0.54	0.19	0.74	0.30
RF	0.61	0.19	0.55	0.28
CatBoost	0.41	0.16	0.81	0.27

Table 71: Performance metrics on the validation set using all predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.63	0.42	0.60	0.50
Linear SVC	0.66	0.44	0.46	0.45
DT	0.59	0.40	0.72	0.52
RF	0.62	0.41	0.55	0.47
CatBoost	0.52	0.36	0.76	0.49

Table 72: Performance metrics on the testing set using all predictors

iv) October

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.61	0.19	0.59	0.29
Linear SVC	0.62	0.62	0.60	0.61
DT	0.58	0.72	0.27	0.39
RF	1	1	1	1
CatBoost	0.92	0.94	0.90	0.92

Table 73: Performance metrics on the training set using all predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.53	0.12	0.63	0.21
Linear SVC	0.78	0.15	0.26	0.19
DT	0.82	0.15	0.20	0.17
RF	0.75	0.14	0.31	0.19
CatBoost	0.84	0.18	0.18	0.18

Table 74: Performance metrics on the validation set using all predictors

Model	Accuracy	Precision	Recall	F1-Score
Full LR	0.48	0.17	0.67	0.27
Linear SVC	0.84	0.22	0.05	0.08
DT	0.82	0.21	0.08	0.12
RF	0.79	0.21	0.17	0.19
CatBoost	0.83	0.29	0.10	0.15

Table 75: Performance metrics on the testing set using all predictors

8.2 Feature Importance Graphs

a) Illinois - Weather

- Feature Importance for Illinois January - Weather Data:

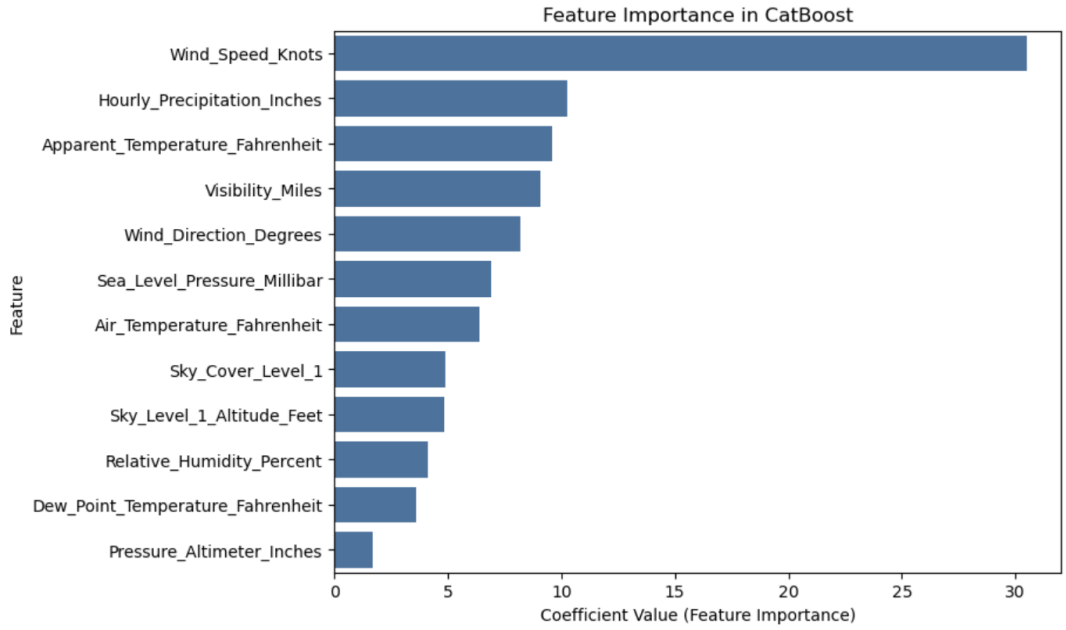


Figure 18: Feature Importance for Illinois January - Weather Data (Graph 1)

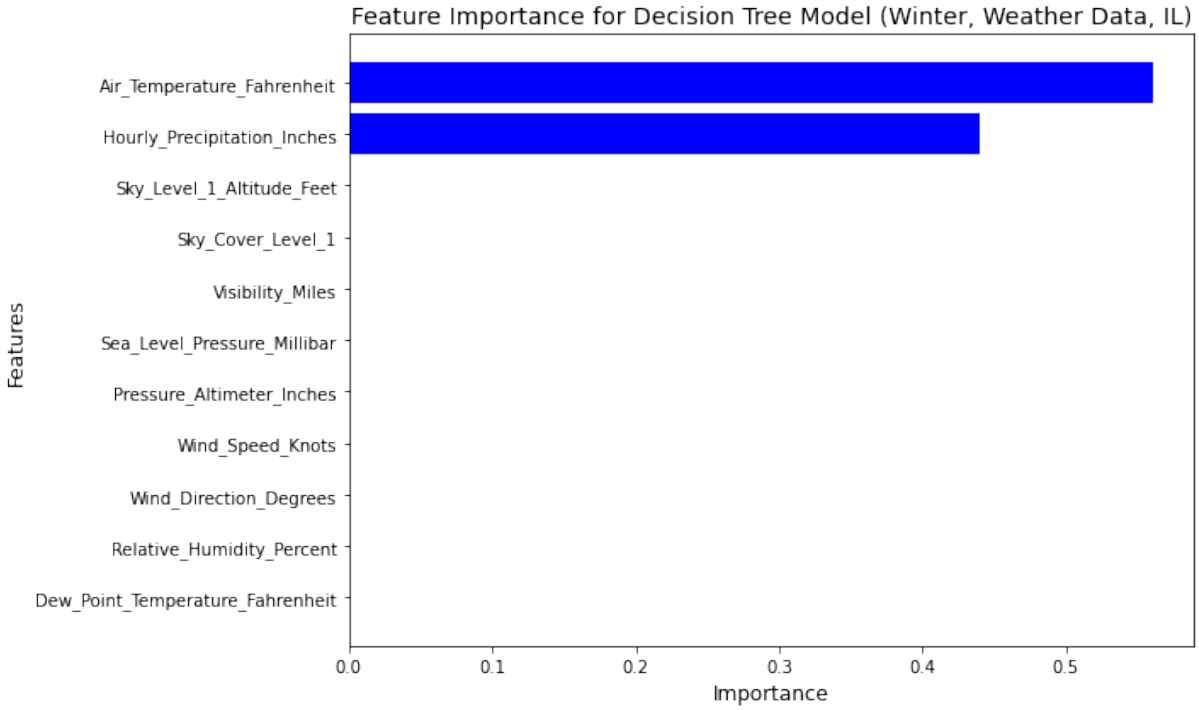


Figure 19: Feature Importance for Illinois January - Weather Data (Graph 2)

b) April

- Feature Importance for Illinois April - Weather Data:

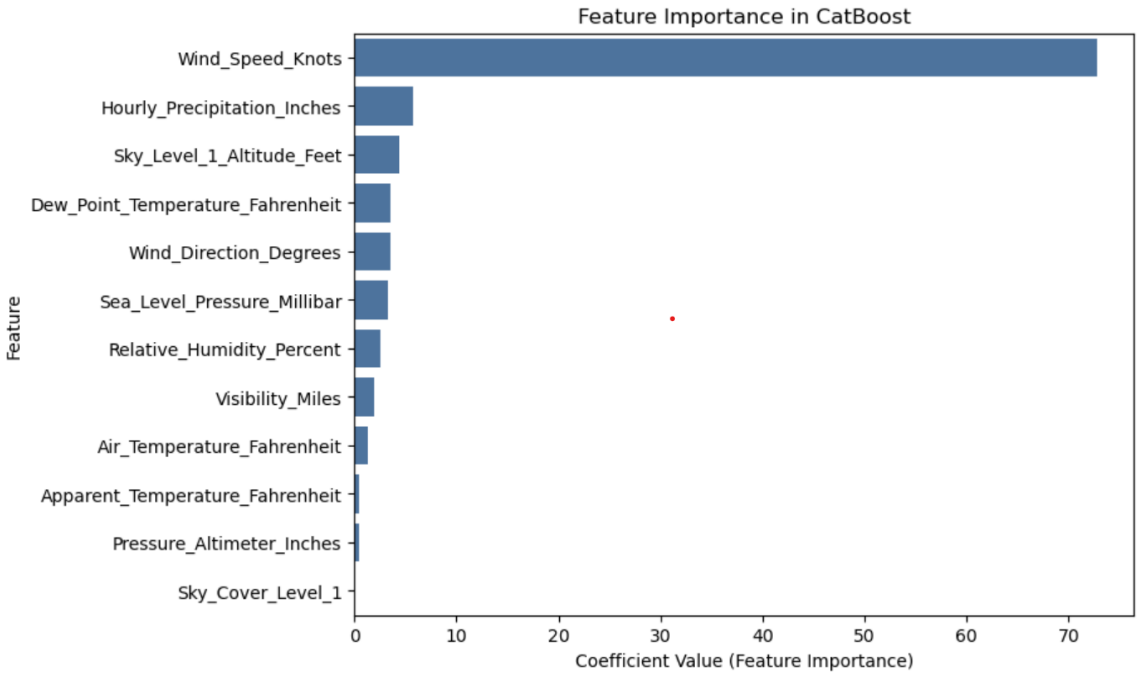


Figure 20: Feature Importance for Illinois April - Weather Data

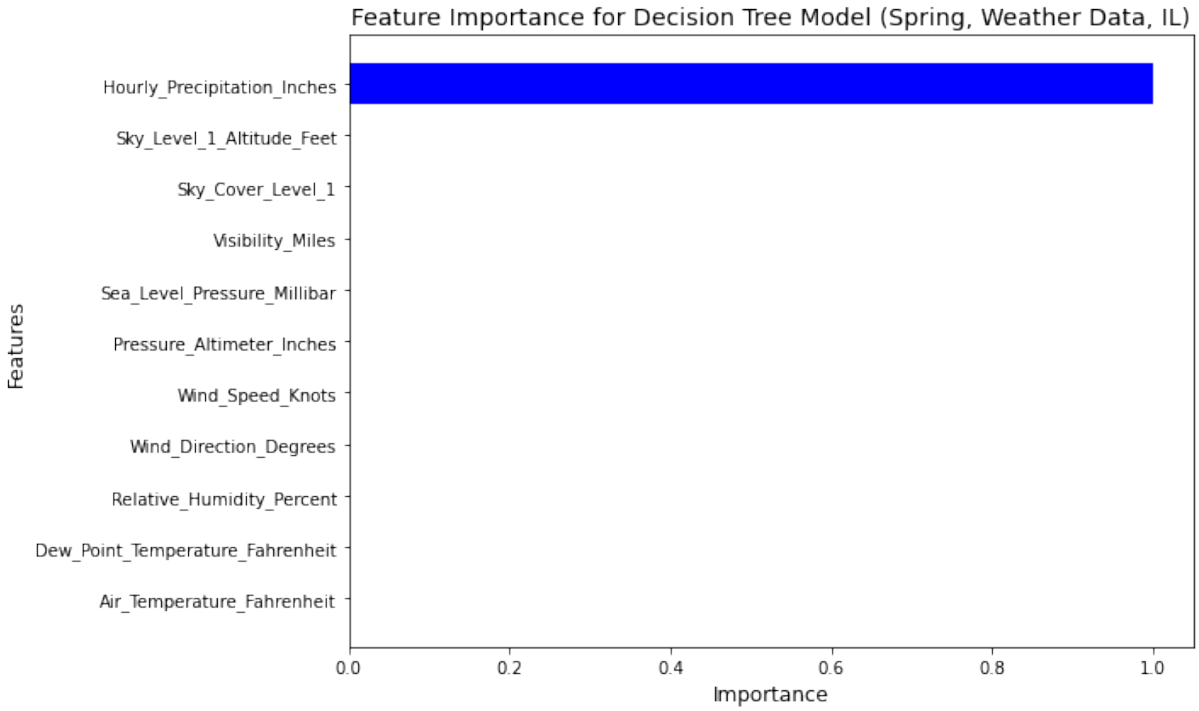


Figure 21: Feature Importance for Illinois April - Weather Data

c) July

- Feature Importance for Illinois July - Weather Data:

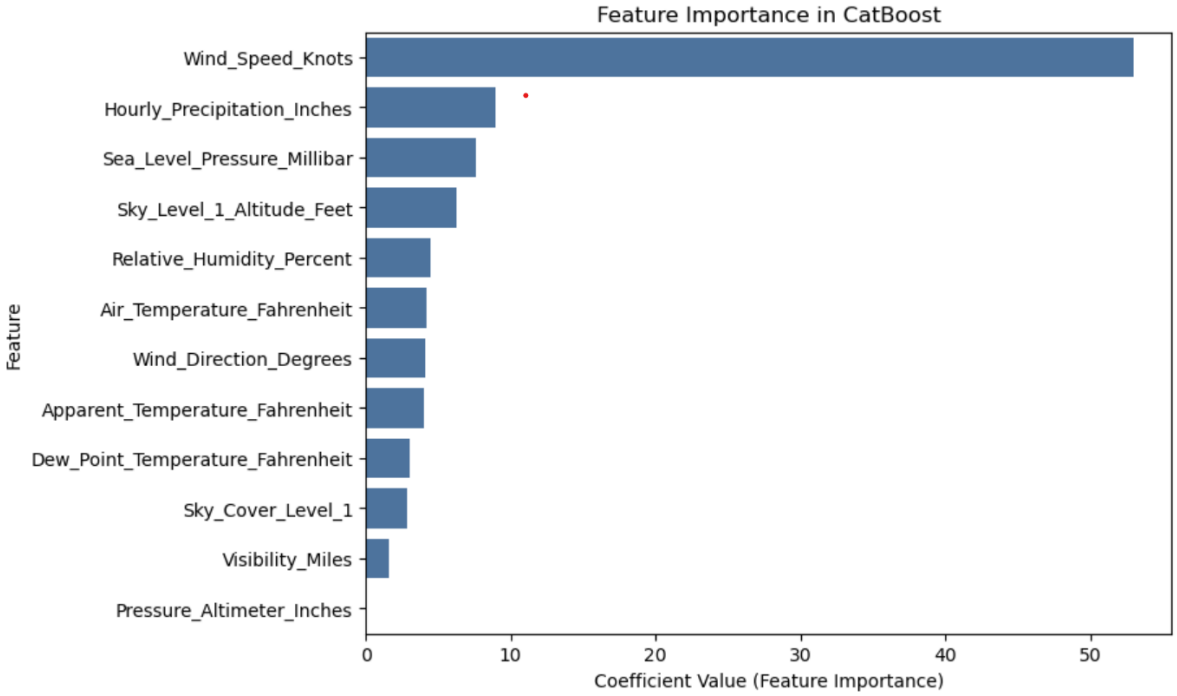


Figure 22: Feature Importance for Illinois July - Weather Data

d) October

- Feature Importance for Illinois October - Weather Data:

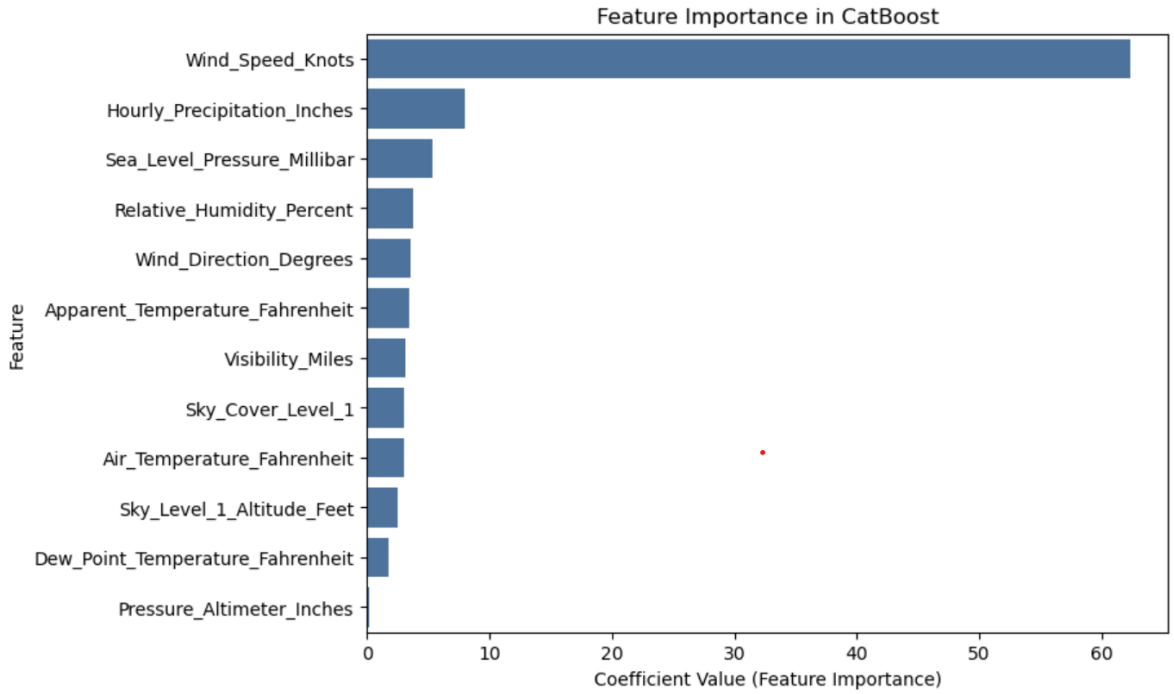


Figure 23: Feature Importance for October - Weather Data

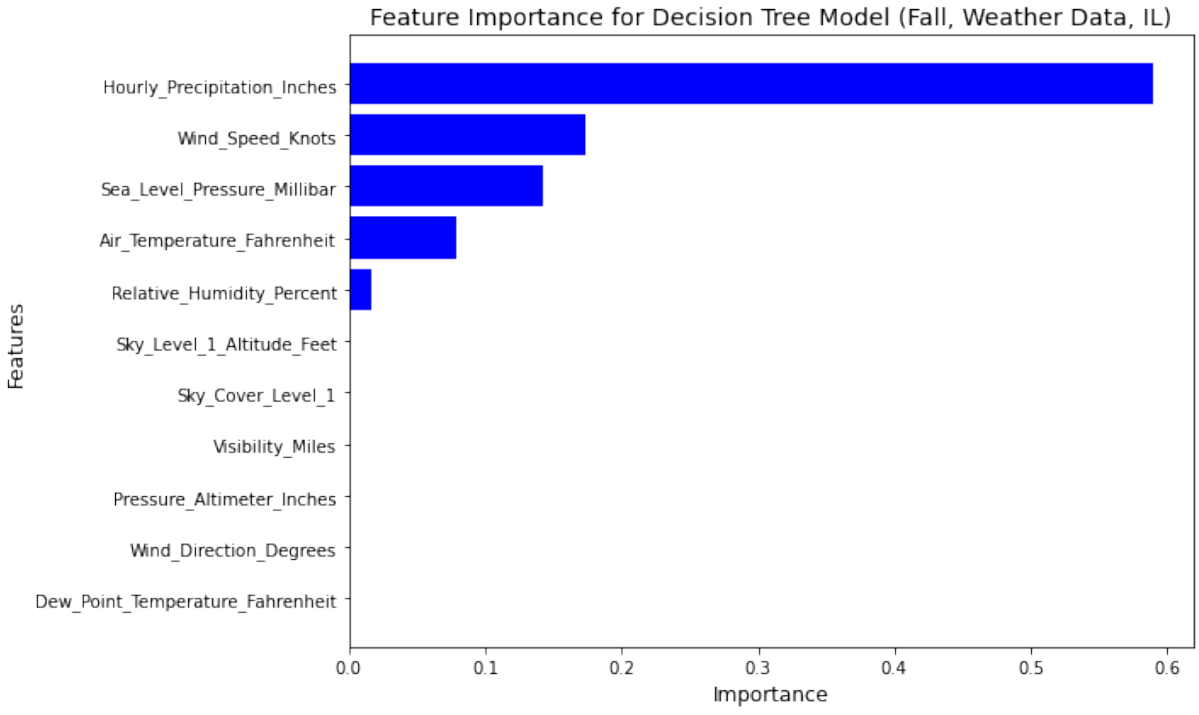


Figure 24: Feature Importance for October - Weather Data

a) **Illinois - Flight**

- Feature Importance for Illinois January - Flight Data:

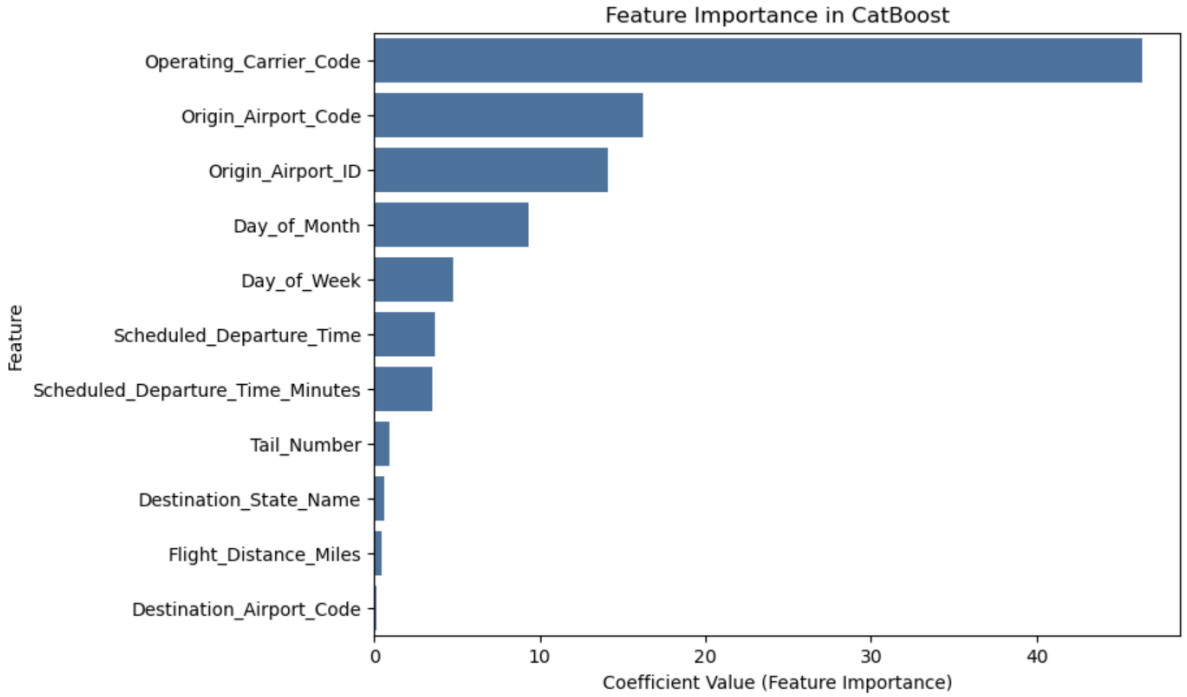


Figure 25: Feature Importance for January - Flight Data

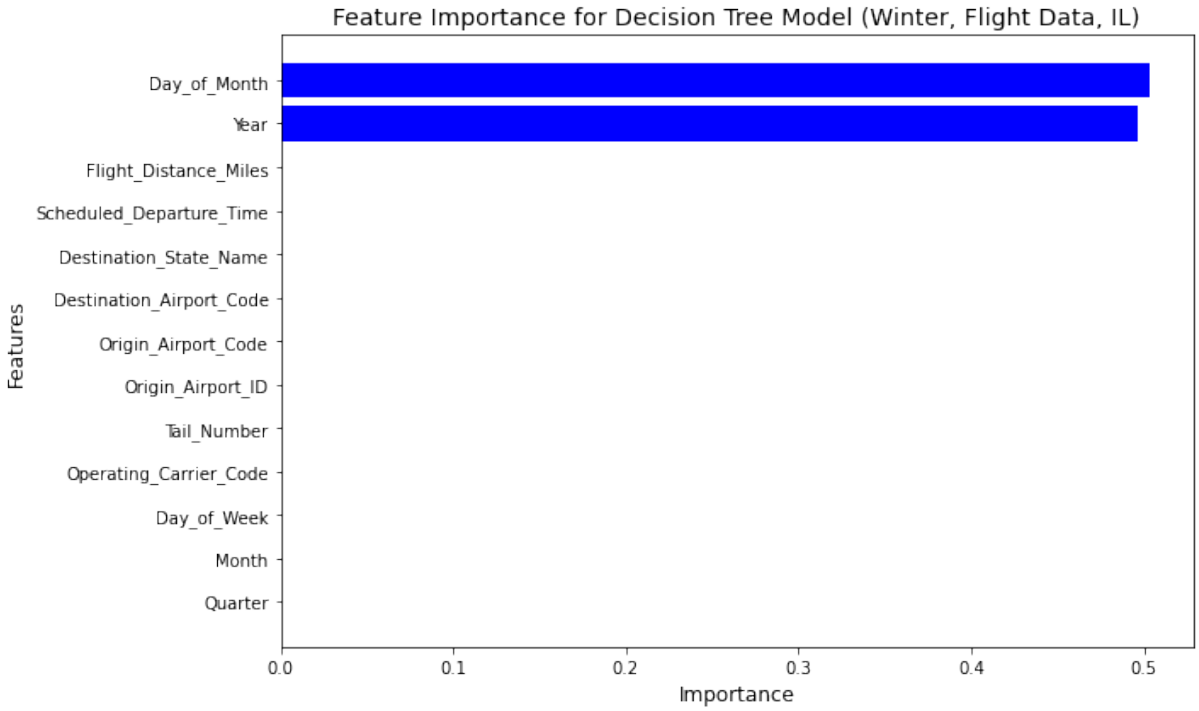


Figure 26: Feature Importance for January - Flight Data

b) April

- Feature Importance for Illinois April - Flight Data:

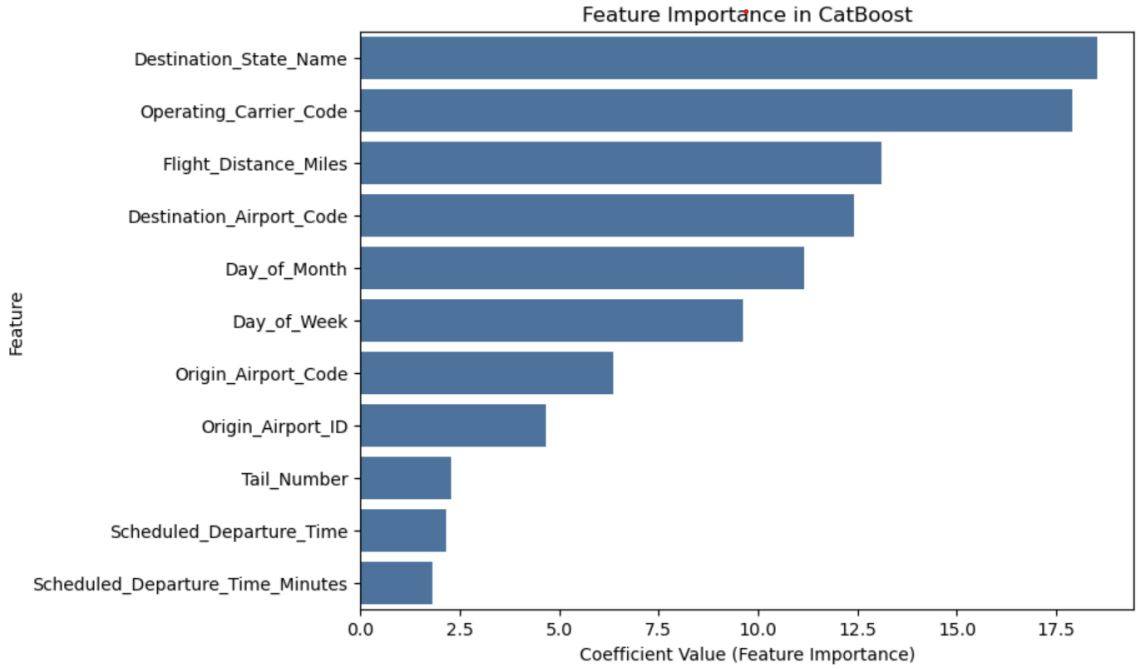


Figure 27: Feature Importance for April - Flight Data

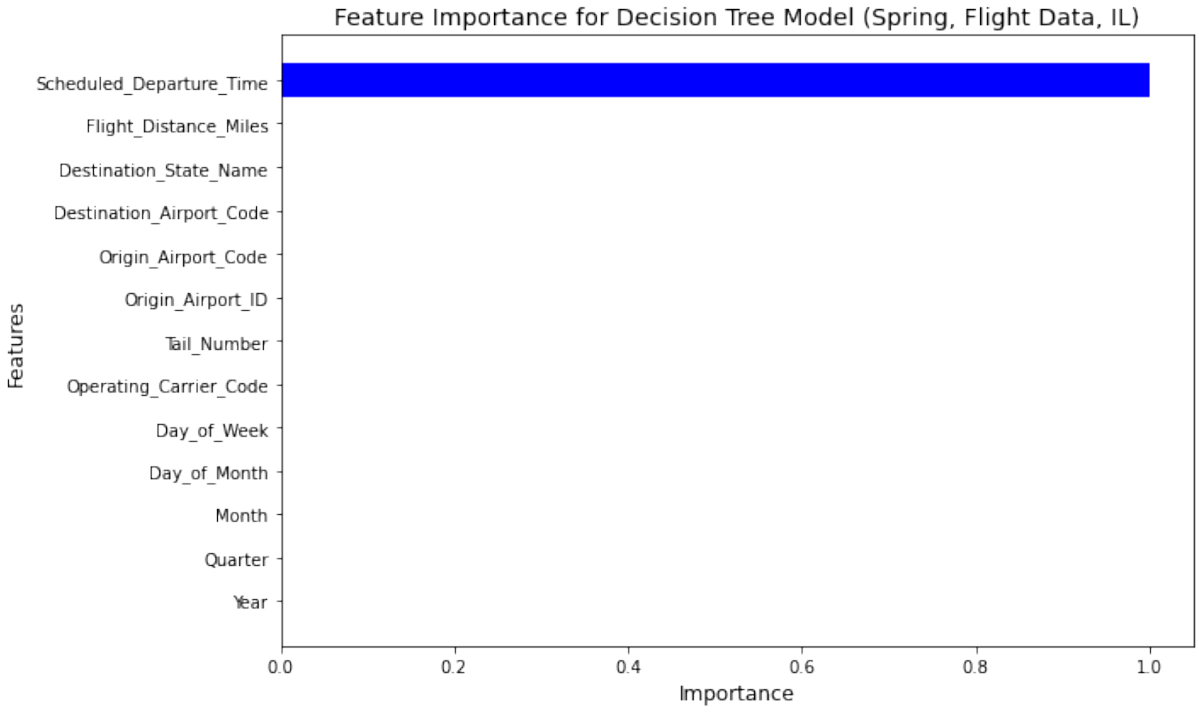


Figure 28: Feature Importance for April - Flight Data

c) July

- Feature Importance for Illinois July - Flight Data:

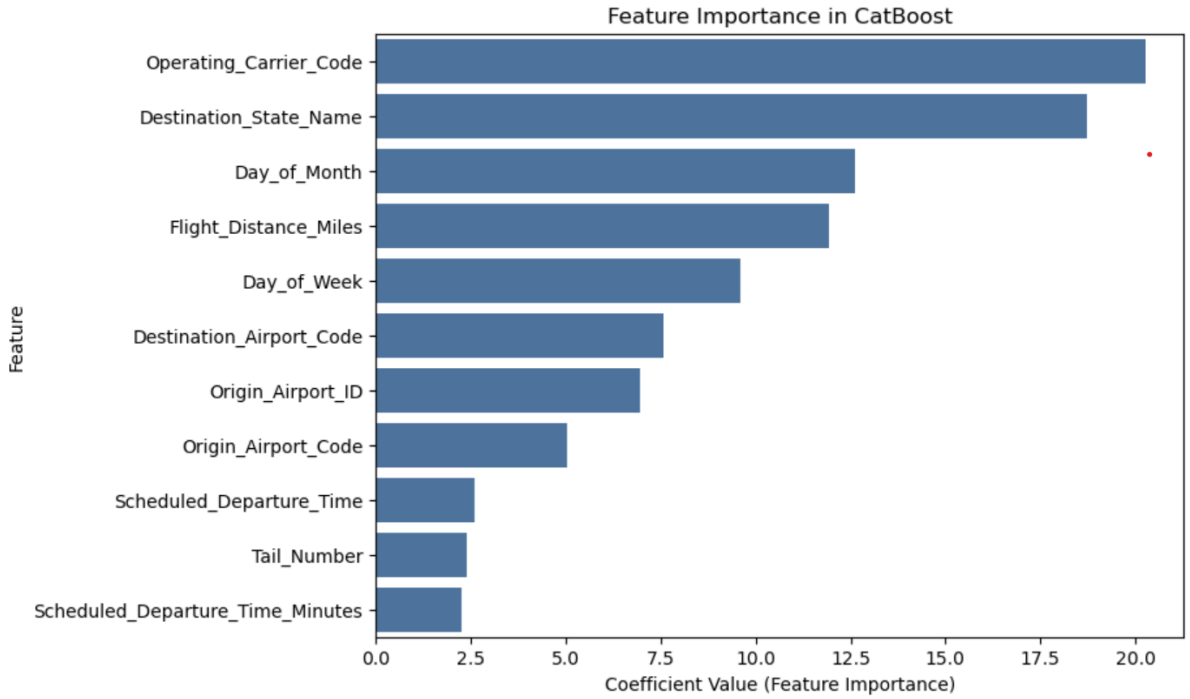


Figure 29: Feature Importance for July - Flight Data

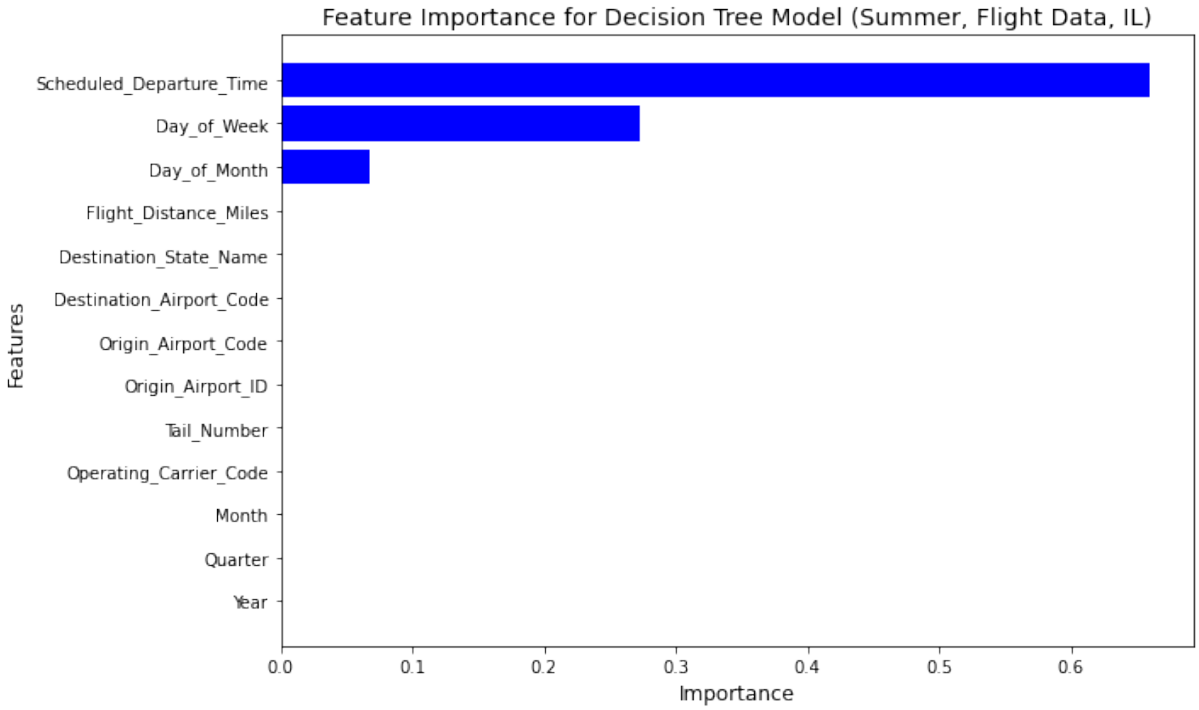


Figure 30: Feature Importance for July - Flight Data

d) October

- Feature Importance for Illinois October - Flight Data:

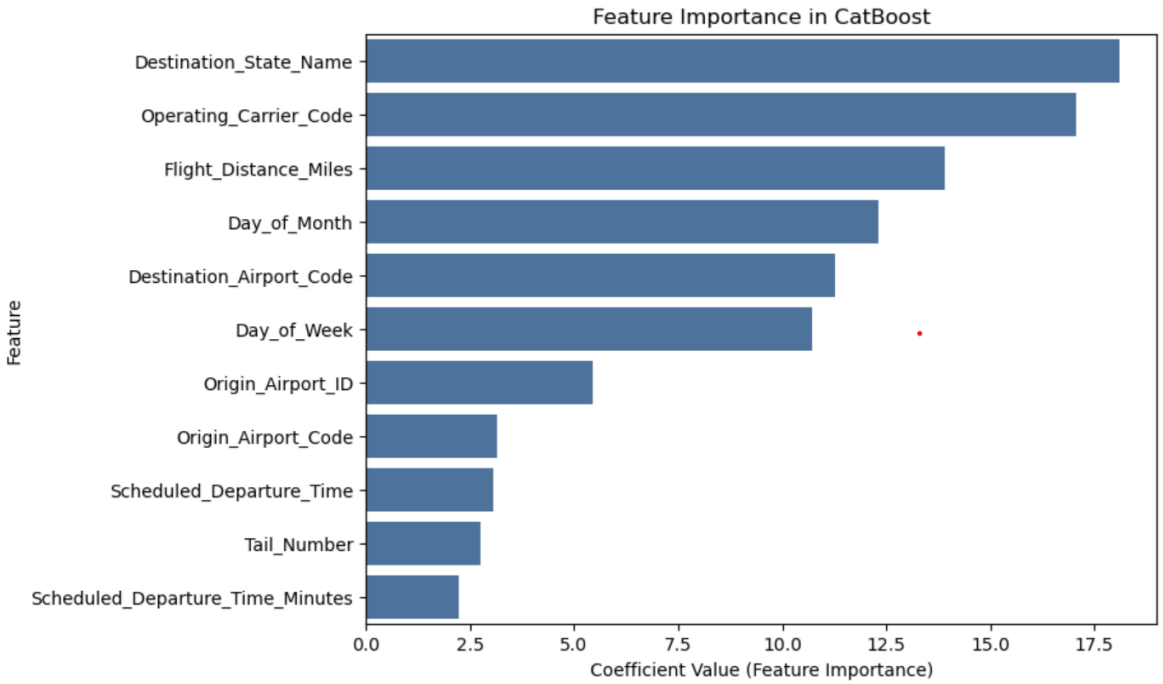


Figure 31: Feature Importance for October - Flight Data

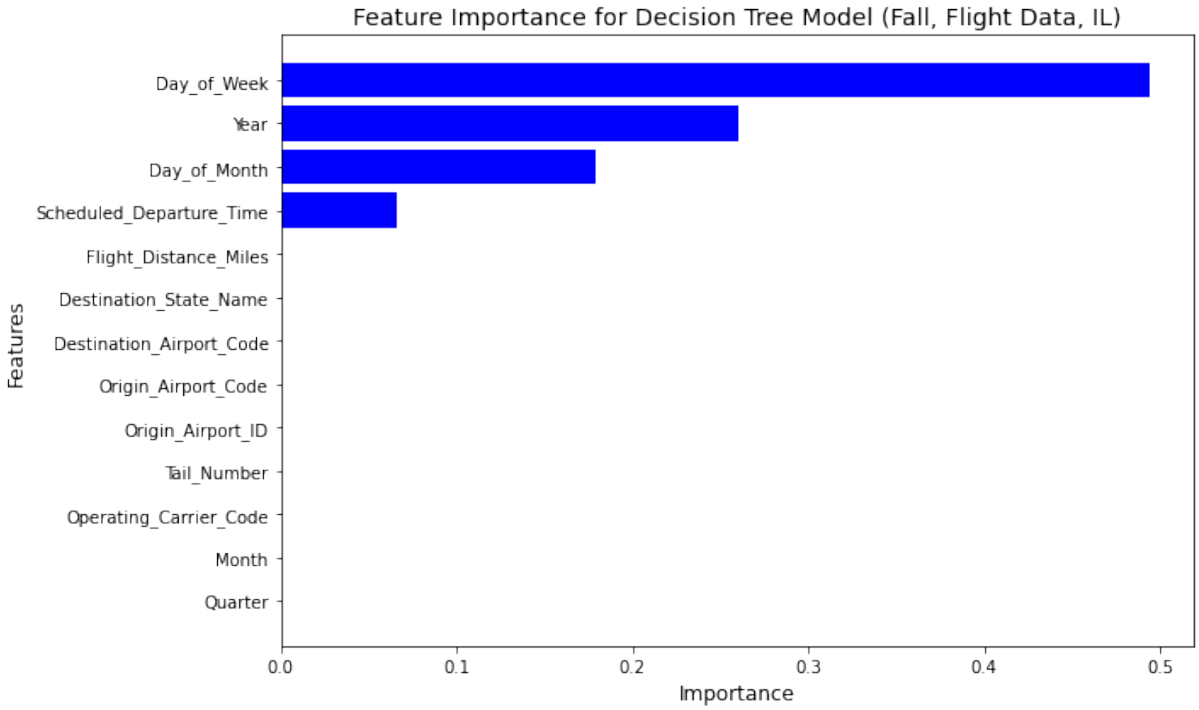


Figure 32: Feature Importance for October - Flight Data

a) **Illinois - Combined Weather and Flight Data**

b) January

- Feature Importance for Illinois January - Combined Weather and Flight Data:

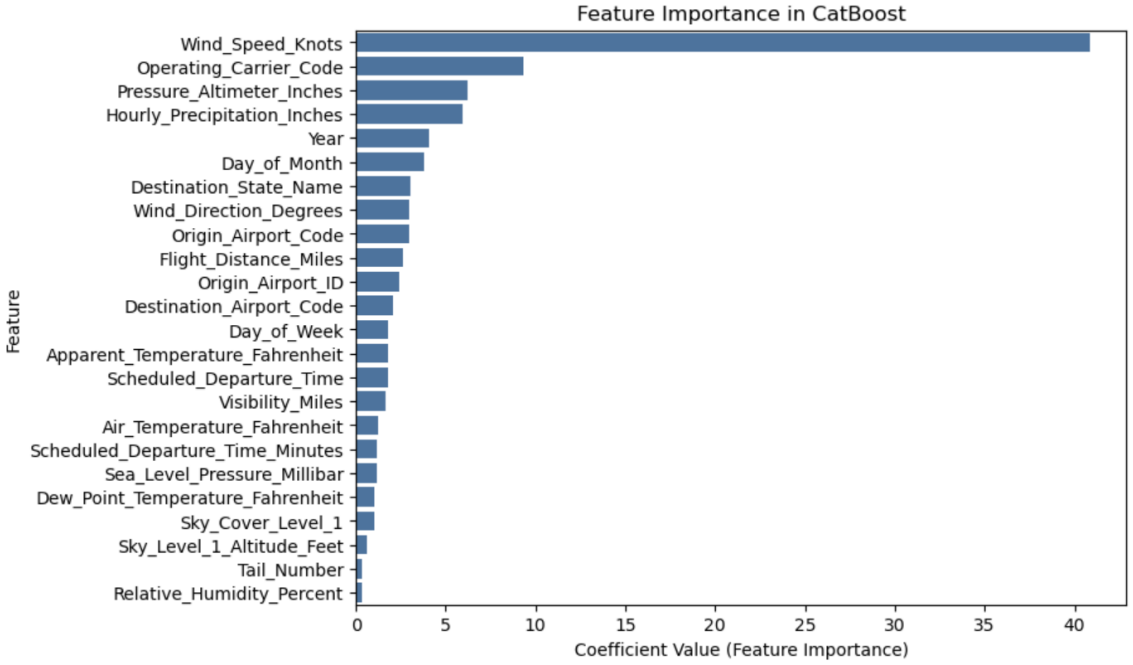


Figure 33: Feature Importance for January - Combined Weather and Flight Data

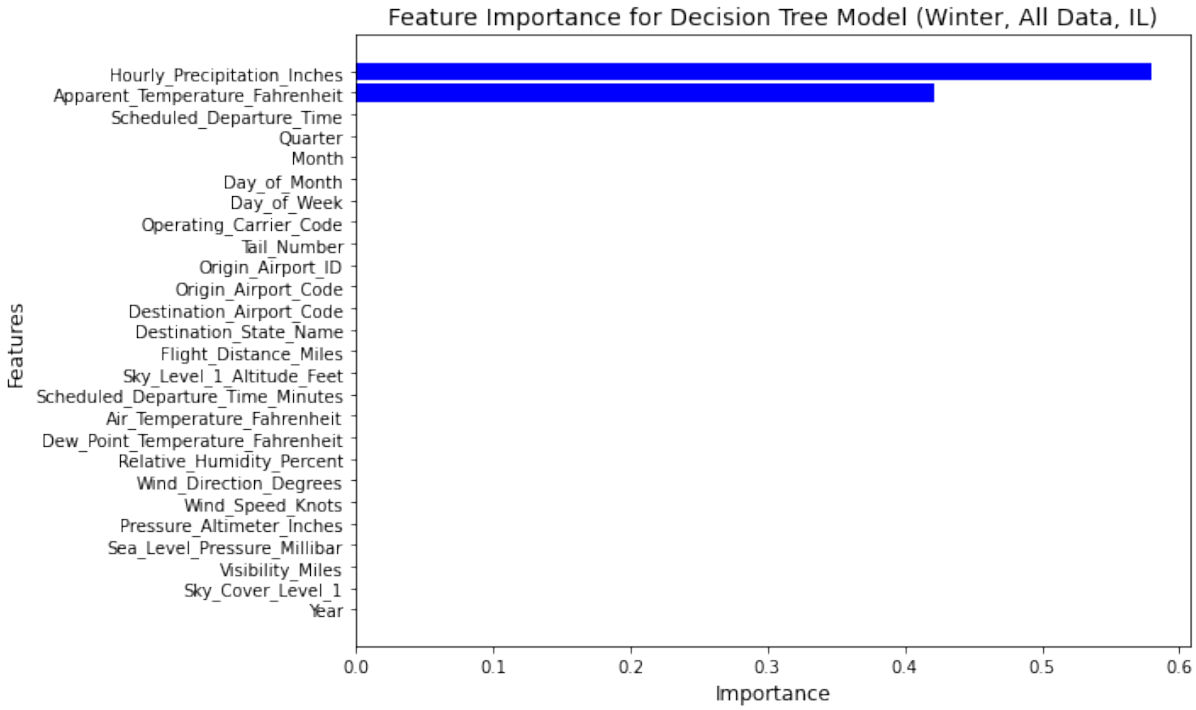


Figure 34: Feature Importance for January - Combined Weather and Flight Data

c) April

- Feature Importance for Illinois April - Combined Weather and Flight Data:

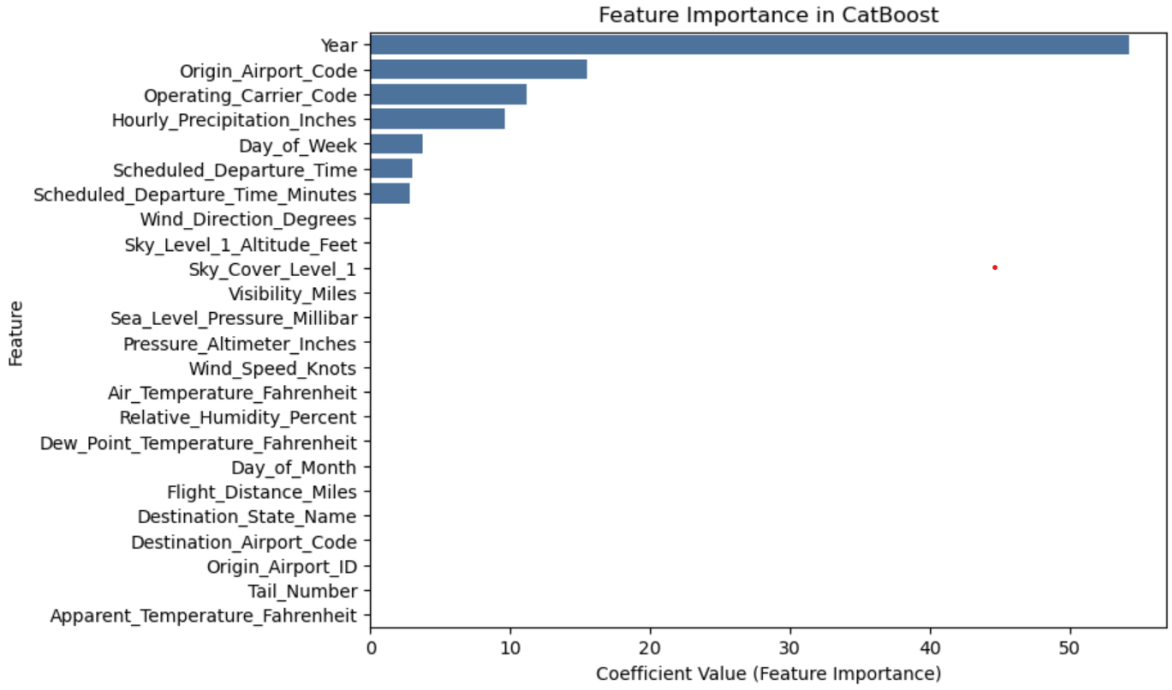


Figure 35: Feature Importance for April - Combined Weather and Flight Data

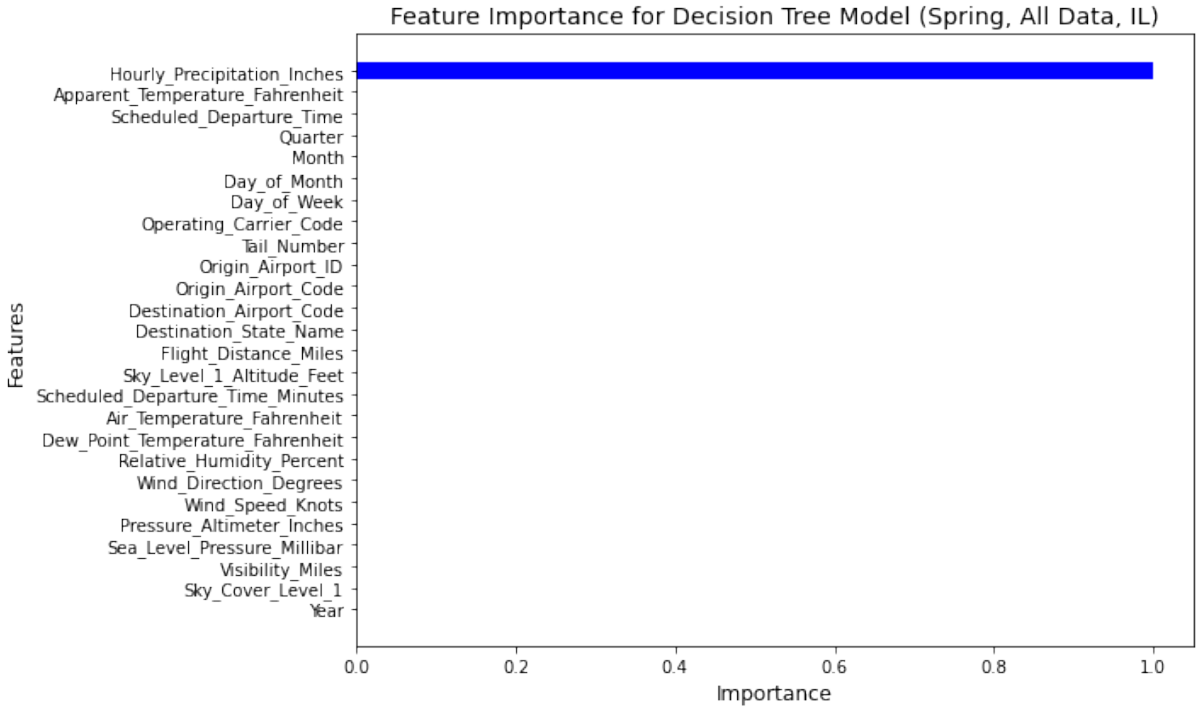


Figure 36: Feature Importance for April - Combined Weather and Flight Data

d) July

- Feature Importance for Illinois July - Combined Weather and Flight Data:

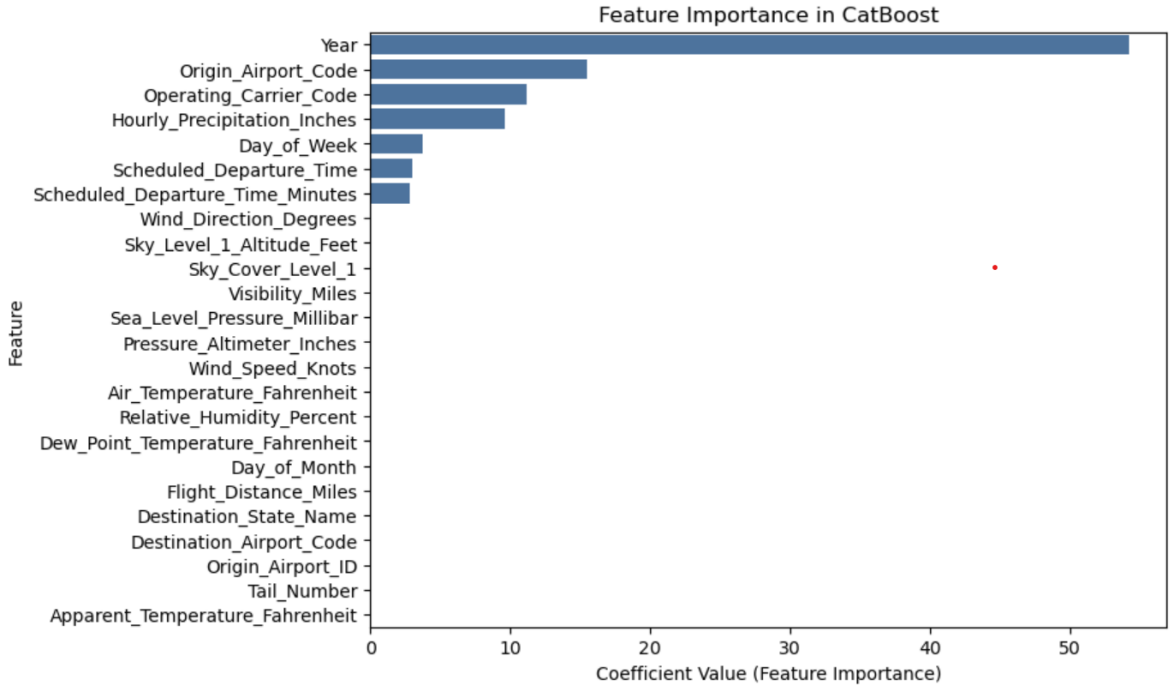


Figure 37: Feature Importance for July - Combined Weather and Flight Data

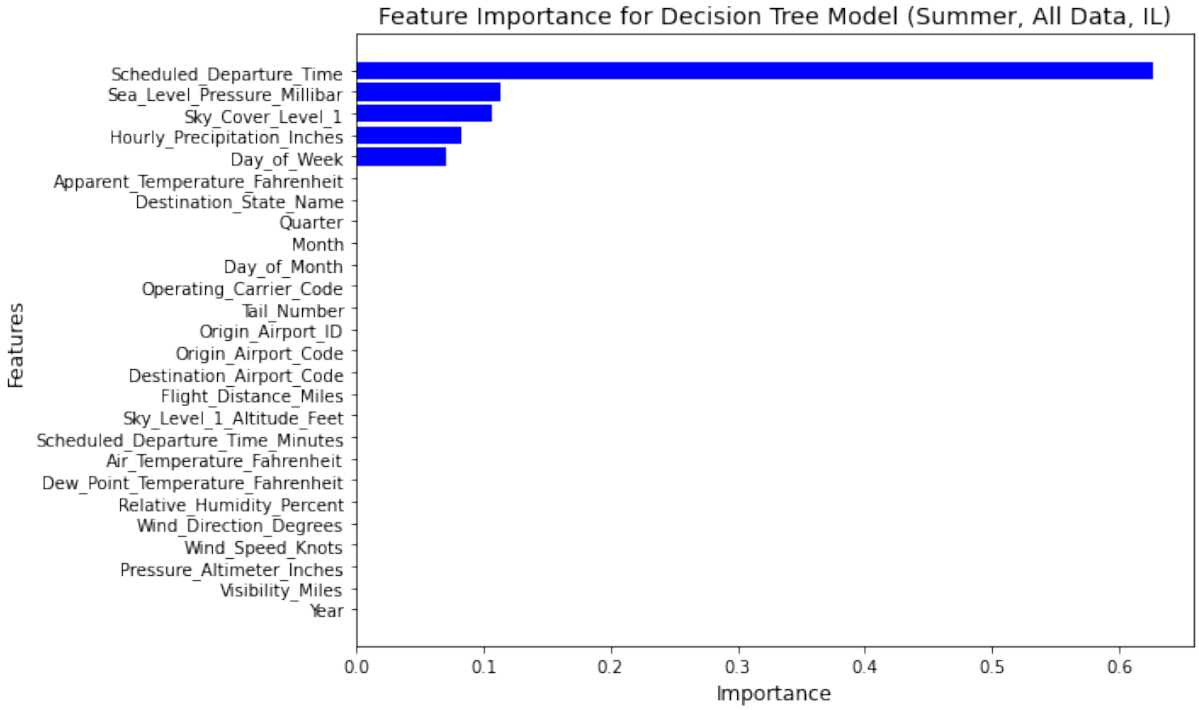


Figure 38: Feature Importance for July - Combined Weather and Flight Data

e) October

- Feature Importance for Illinois October - Combined Weather and Flight Data:

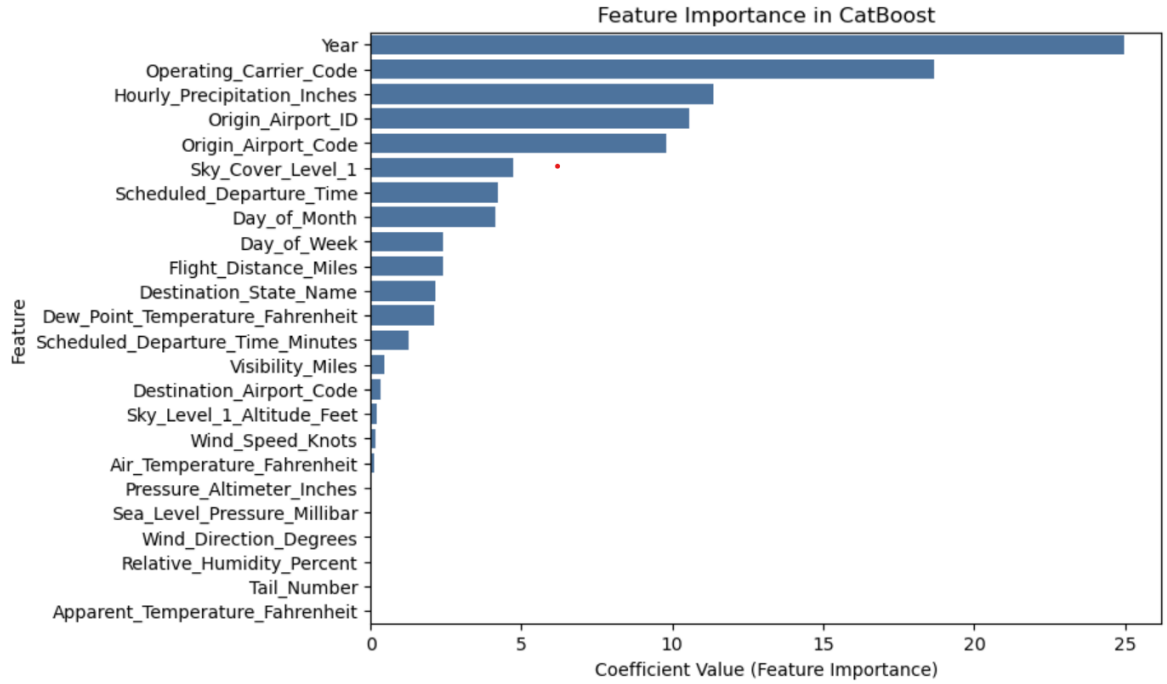


Figure 39: Feature Importance for October - Combined Weather and Flight Data

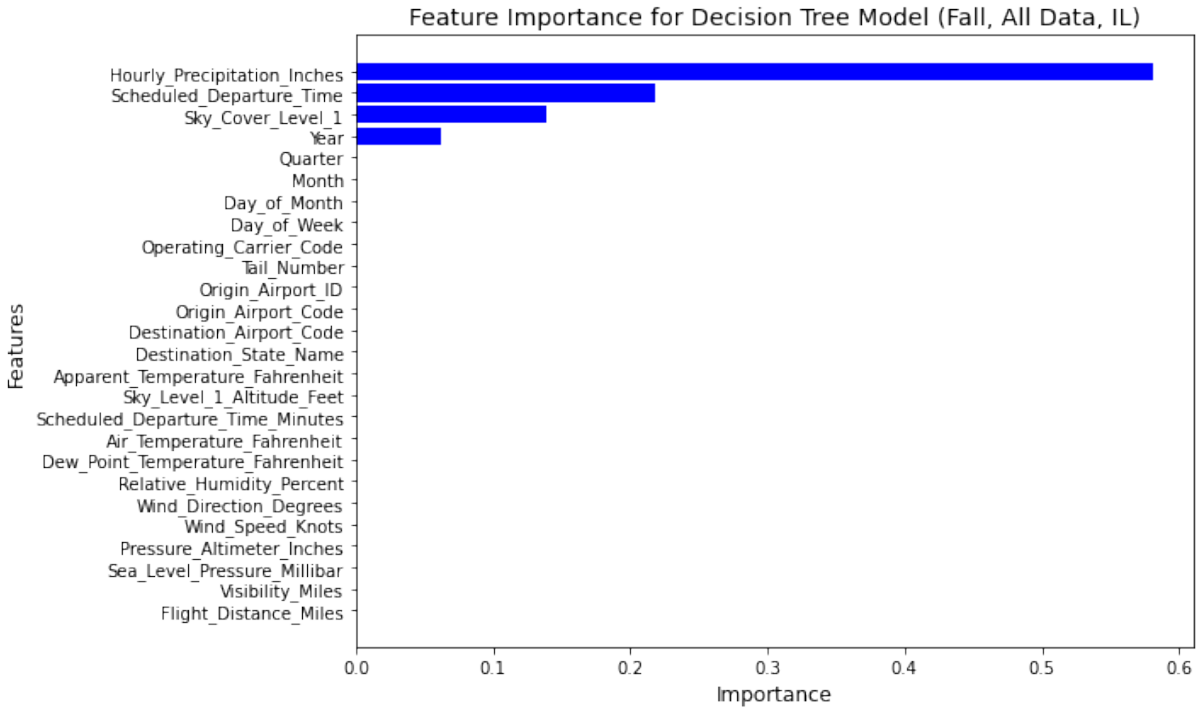


Figure 40: Feature Importance for October - Combined Weather and Flight Data

a) **Georgia - Weather**

b) January

- Feature Importance for Georgia January - Weather Data:

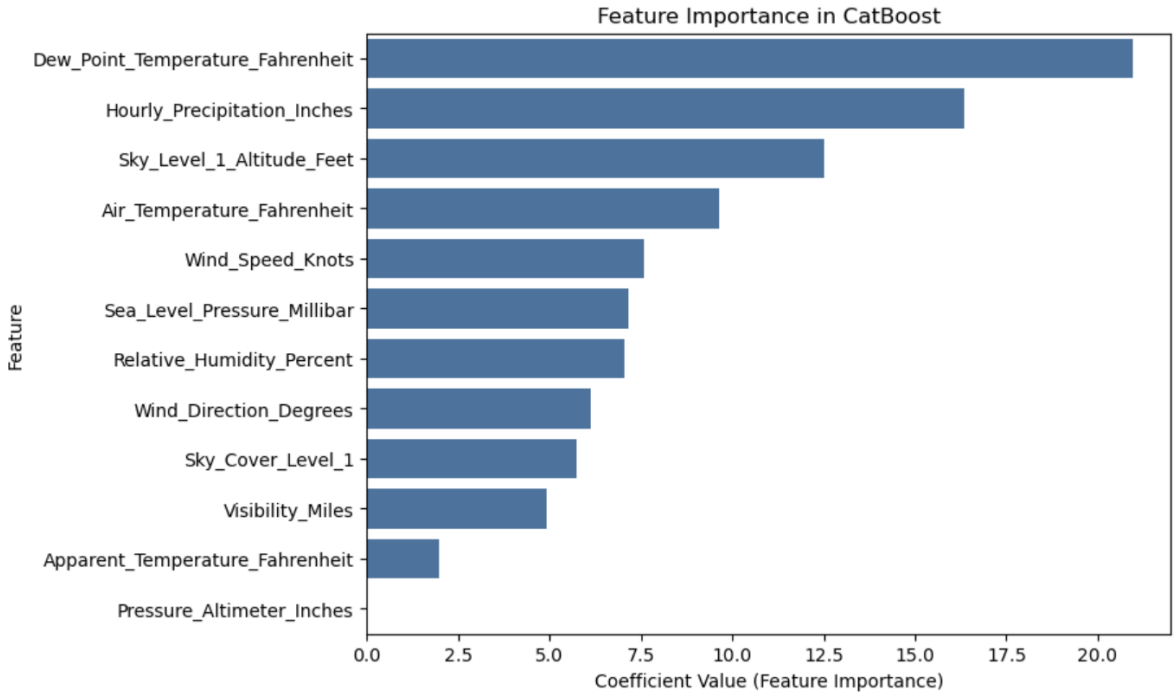


Figure 41: Feature Importance for January - Weather Data

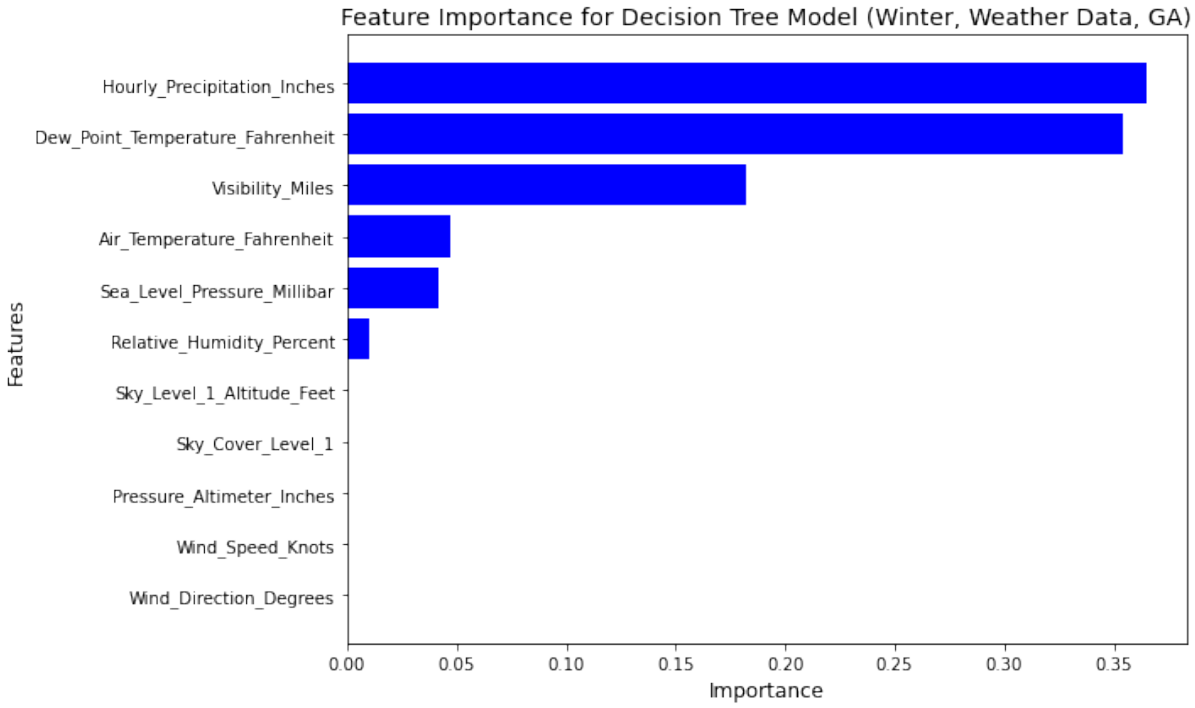


Figure 42: Feature Importance for January - Weather Data

c) April

- Feature Importance for Georgia April - Weather Data:

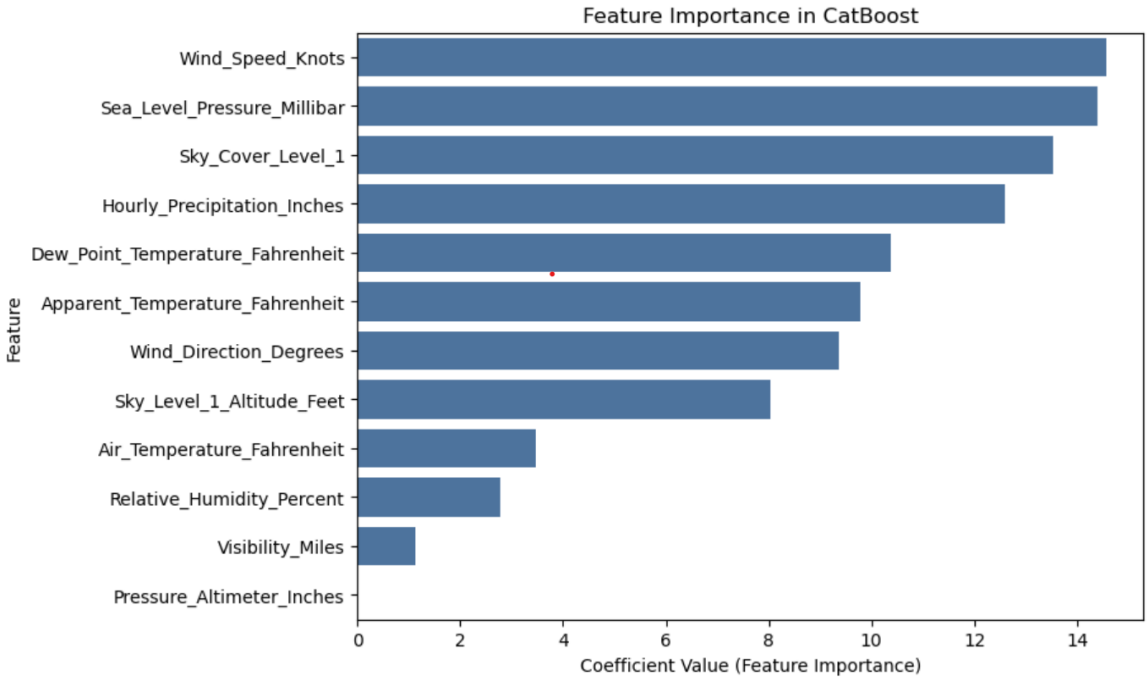


Figure 43: Feature Importance for April - Weather Data

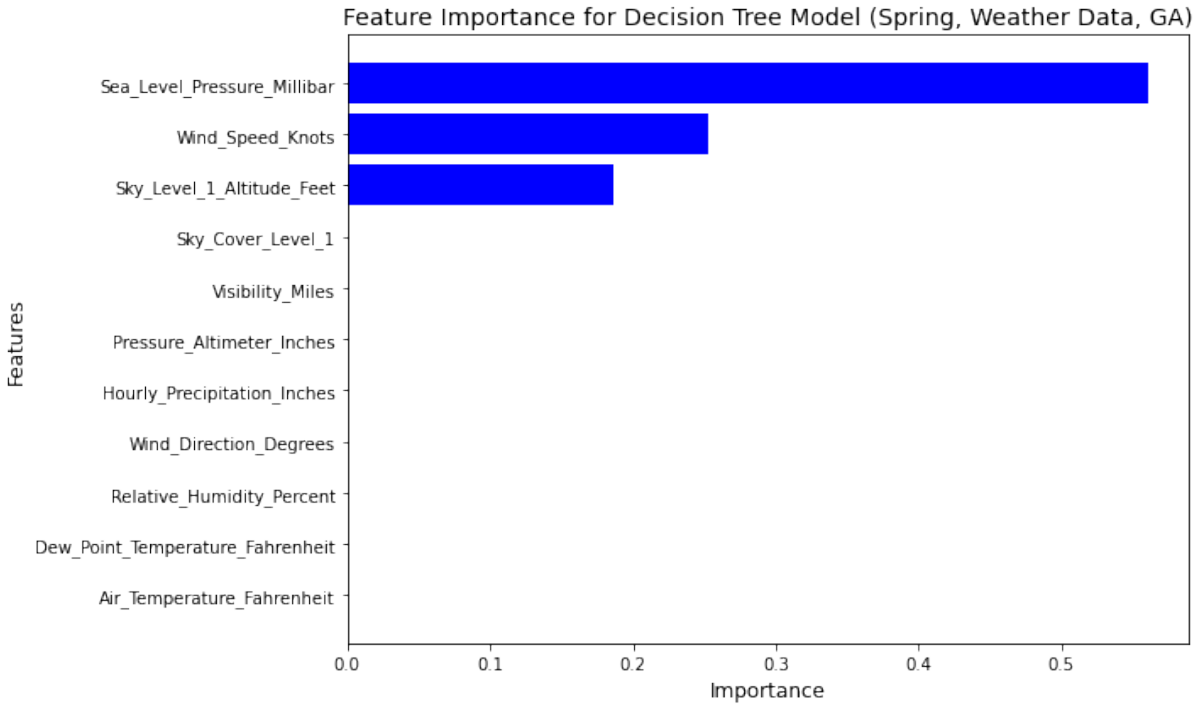


Figure 44: Feature Importance for April - Weather Data

d) July

- Feature Importance for Georgia July - Weather Data:

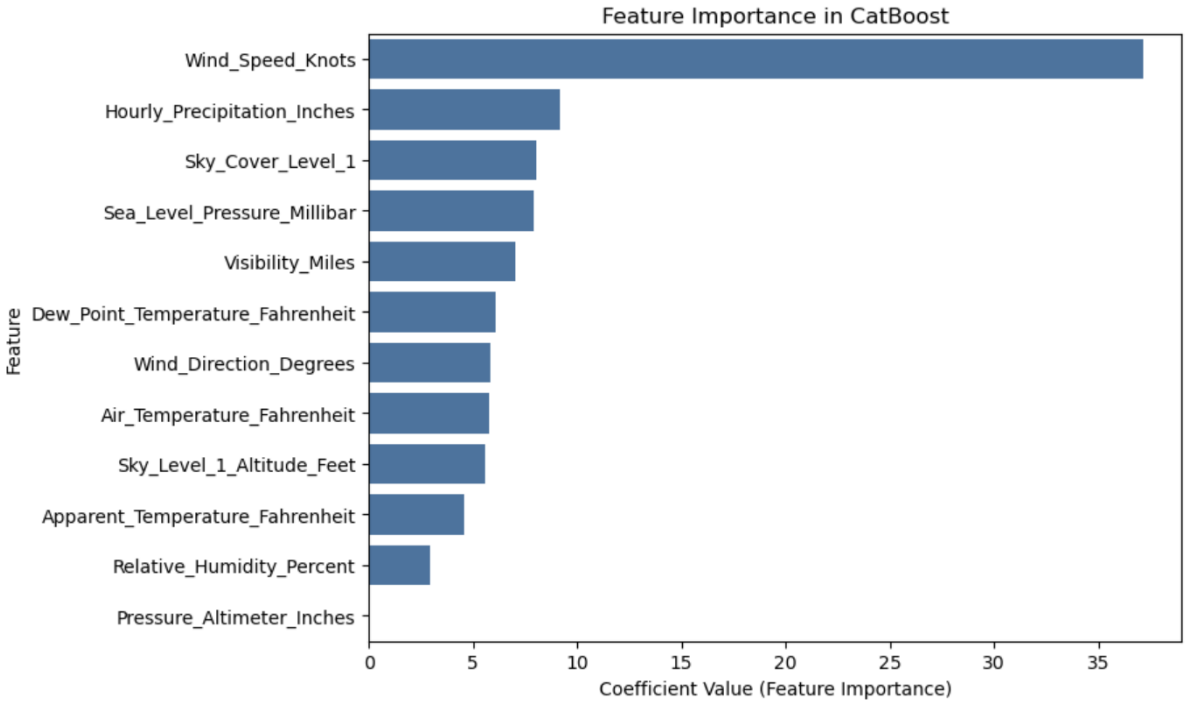


Figure 45: Feature Importance for July - Weather Data

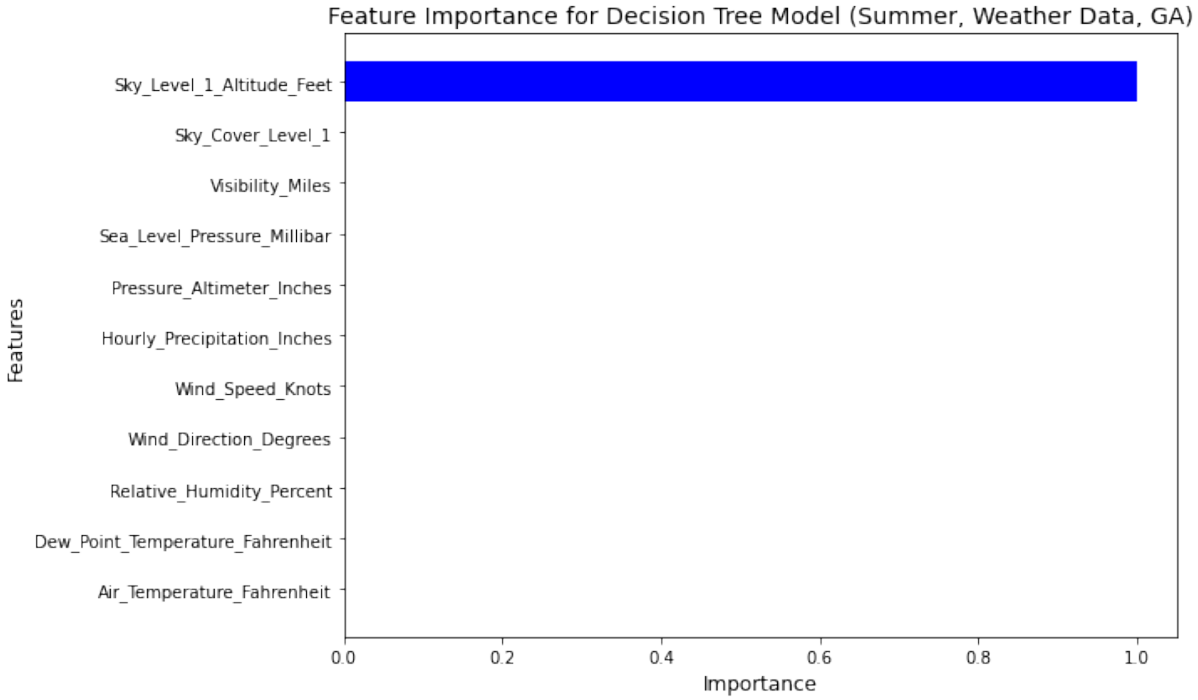


Figure 46: Feature Importance for July - Weather Data

e) October

- Feature Importance for Georgia October - Weather Data:

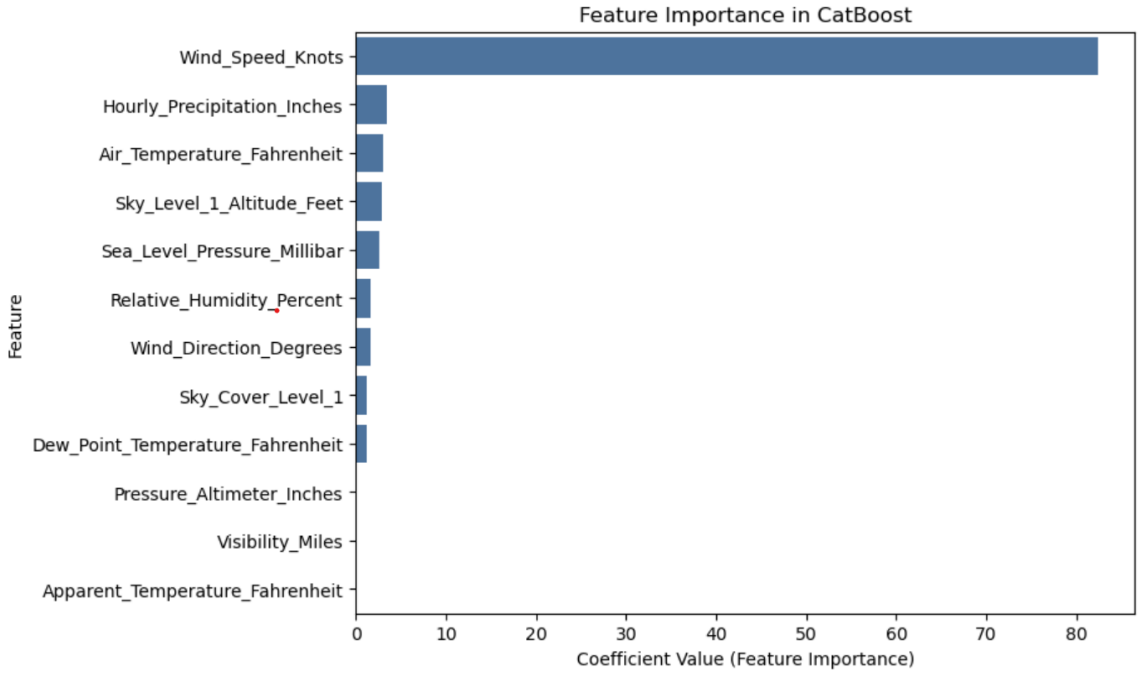


Figure 47: Feature Importance for October - Weather Data

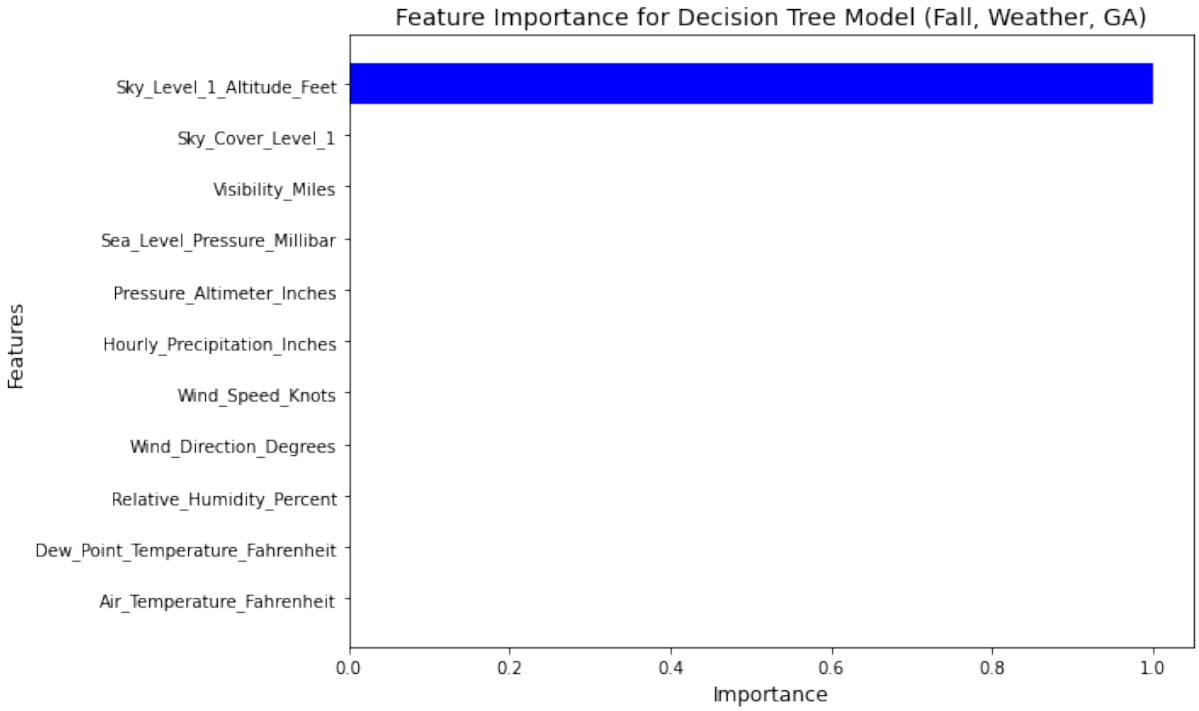


Figure 48: Feature Importance for October - Weather Data

a) **Georgia - Flight**

b) January

- Feature Importance for Georgia January - Flight Data:

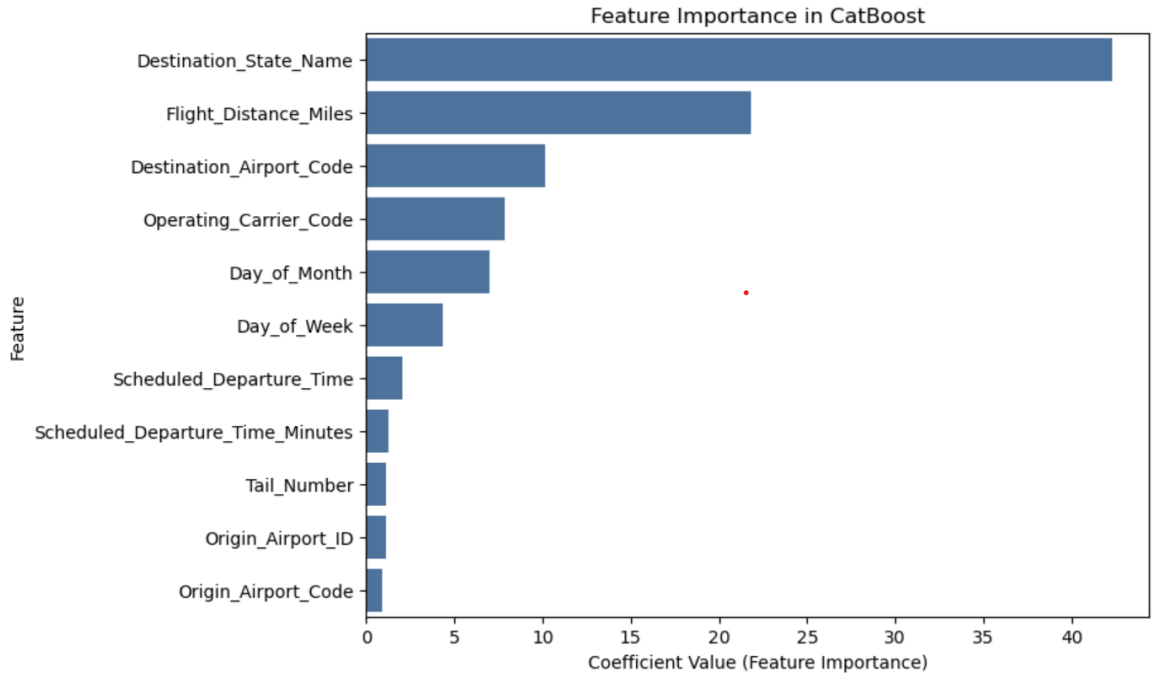


Figure 49: Feature Importance for January - Flight Data

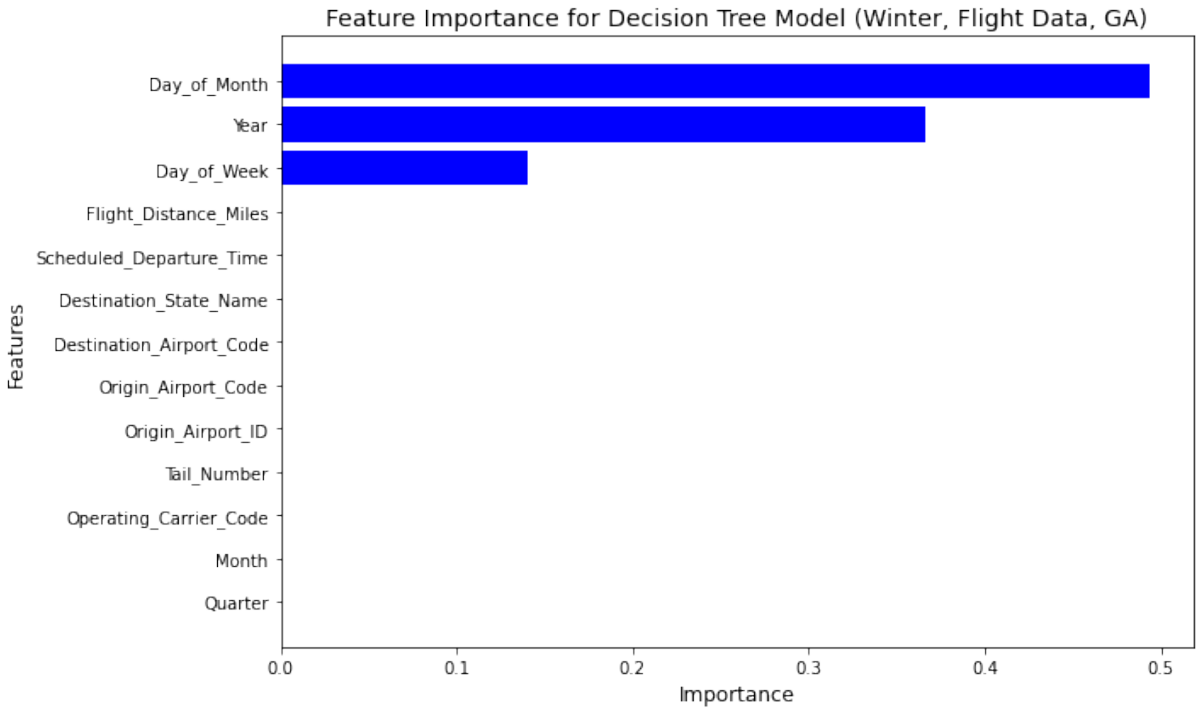


Figure 50: Feature Importance for January - Flight Data

c) April

- Feature Importance for Georgia April - Flight Data:

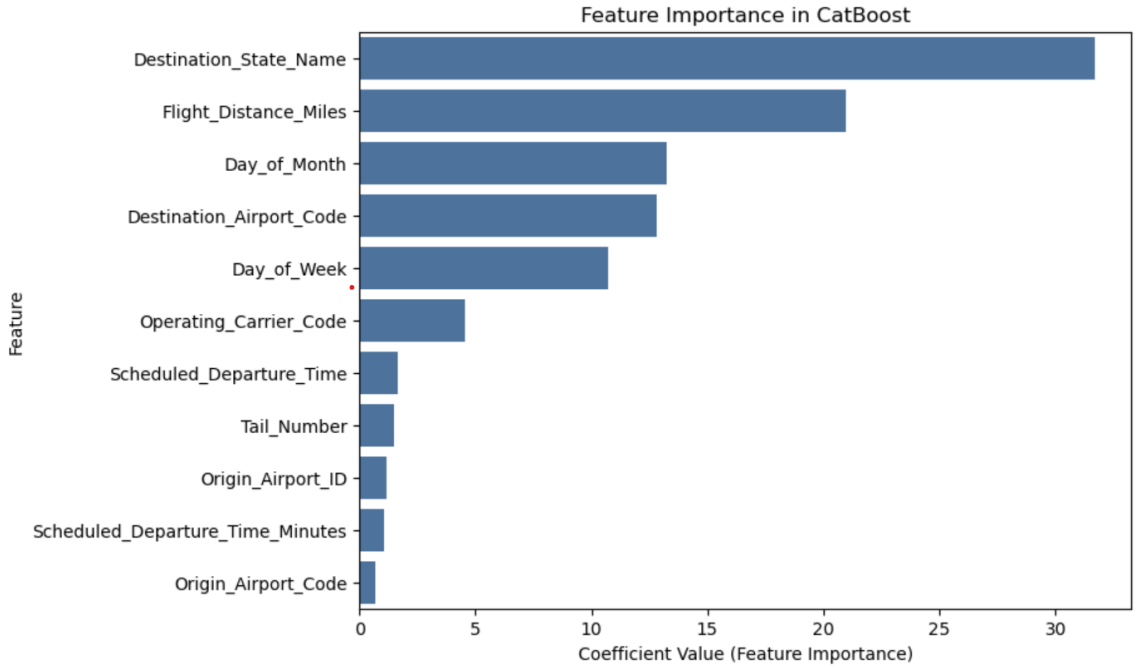


Figure 51: Feature Importance for April - Flight Data

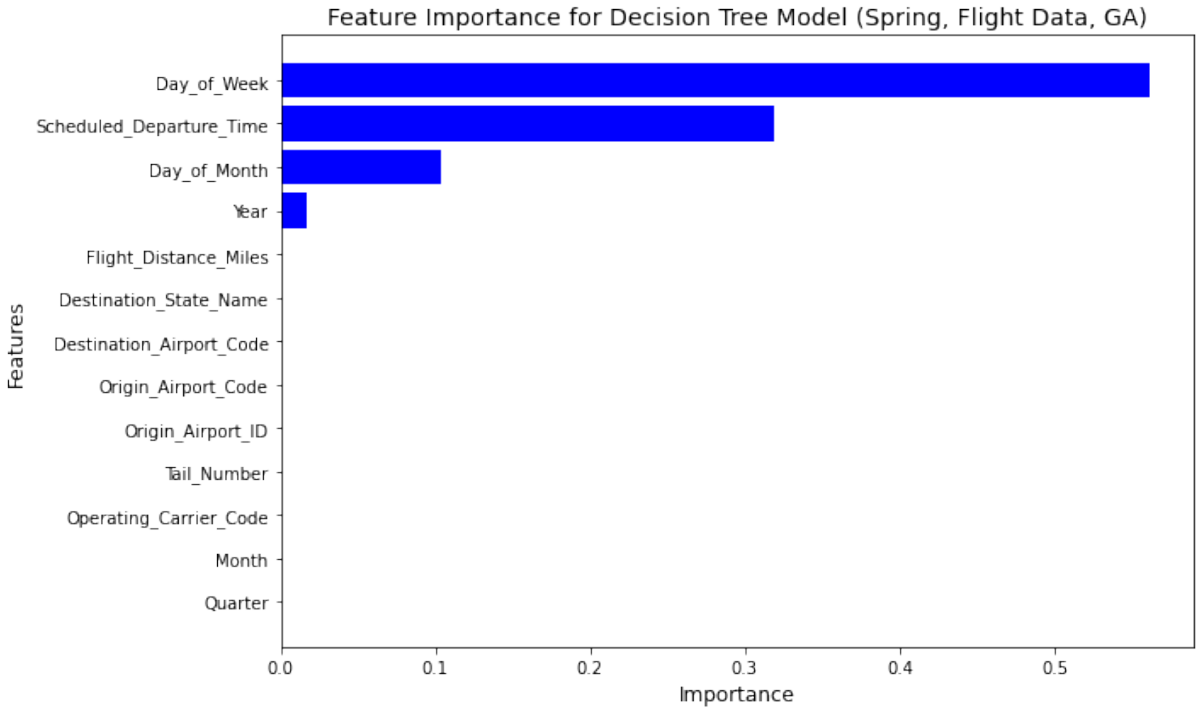


Figure 52: Feature Importance for April - Flight Data

d) July

- Feature Importance for Georgia July - Flight Data:

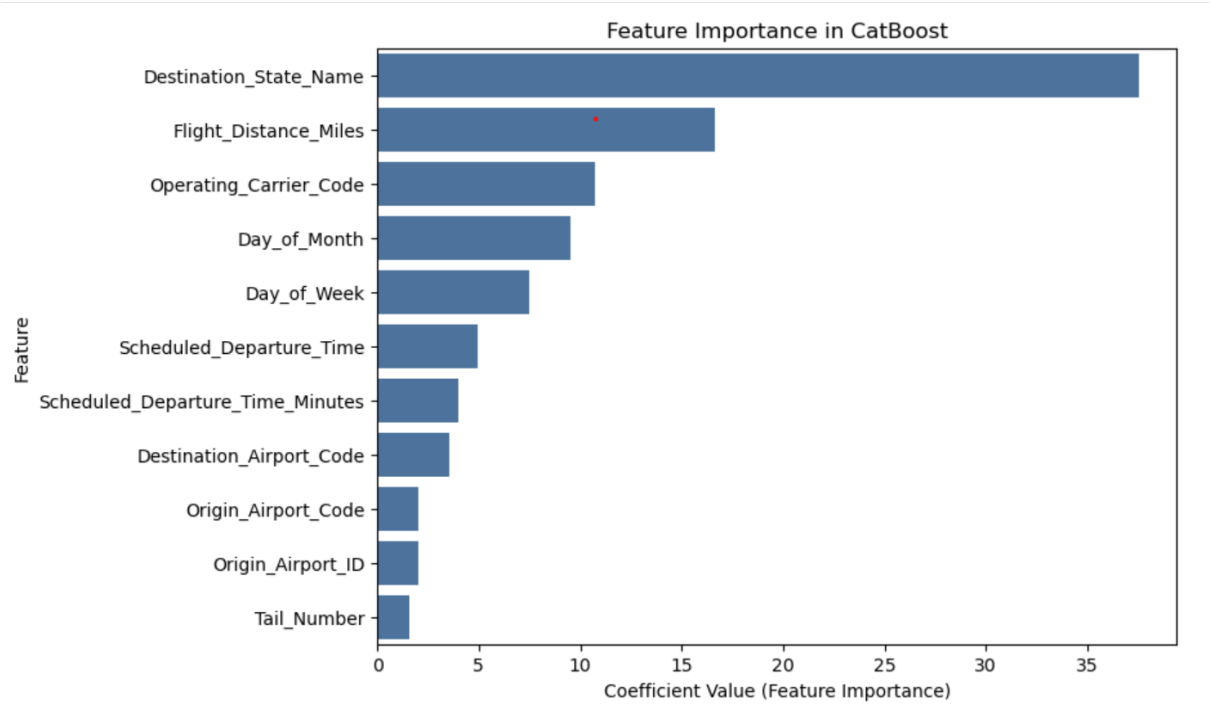


Figure 53: Feature Importance for July - Flight Data

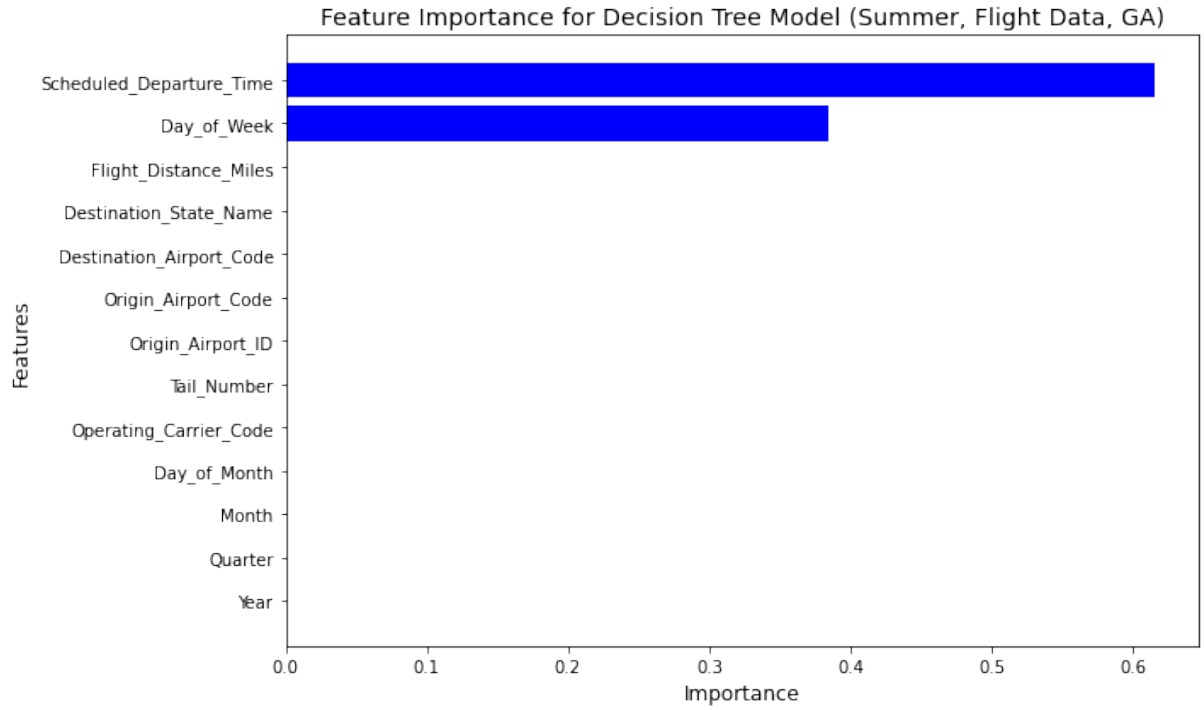


Figure 54: Feature Importance for July - Flight Data

e) October

- Feature Importance for Georgia October - Flight Data:

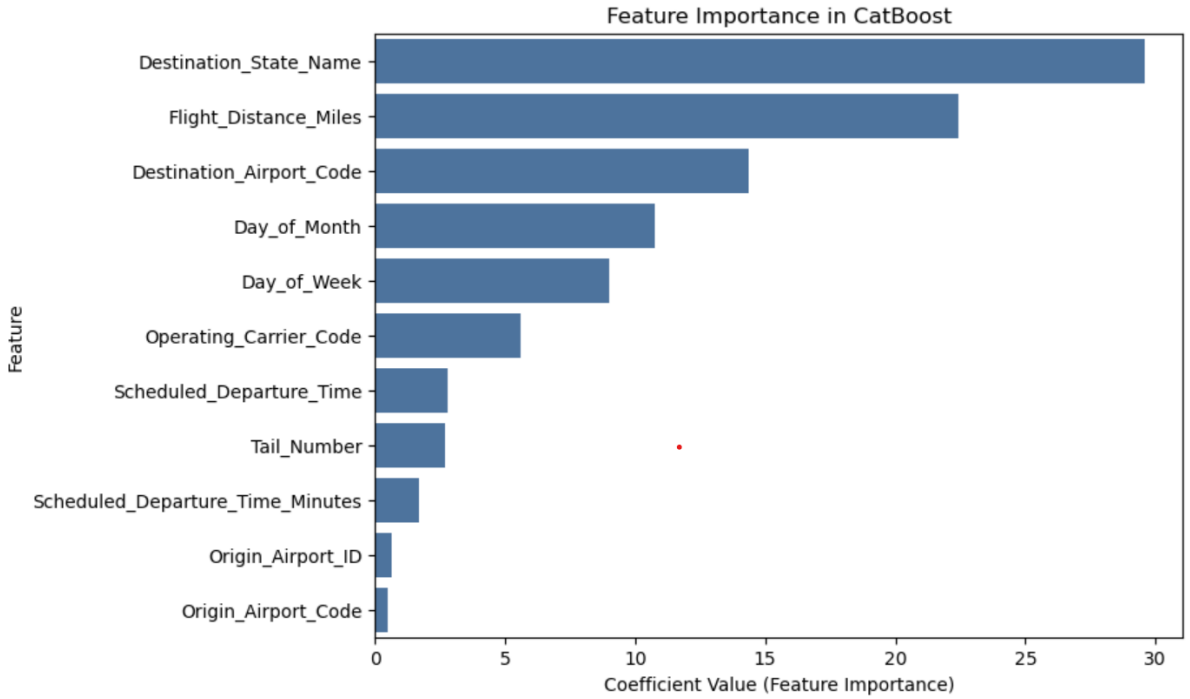


Figure 55: Feature Importance for October - Flight Data

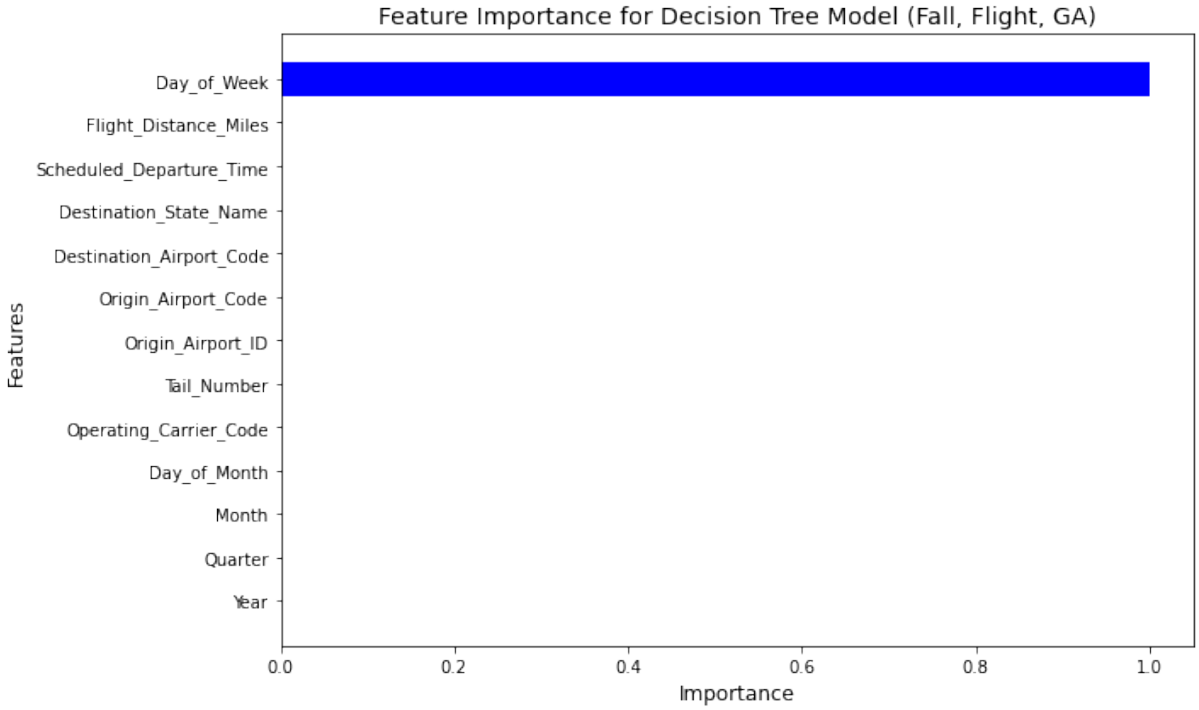


Figure 56: Feature Importance for October - Flight Data

a) **Georgia - Combined Weather and Flight Data**

b) January

- Feature Importance for Georgia January - Combined Weather and Flight Data:

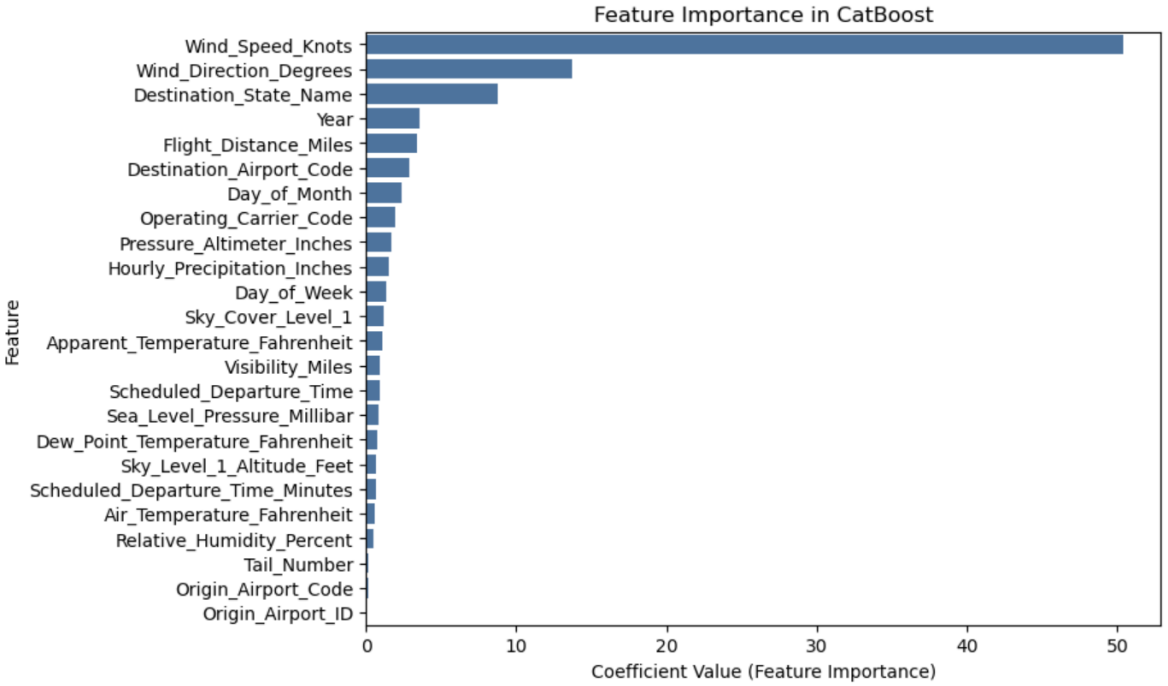


Figure 57: Feature Importance for January - Combined Weather and Flight Data

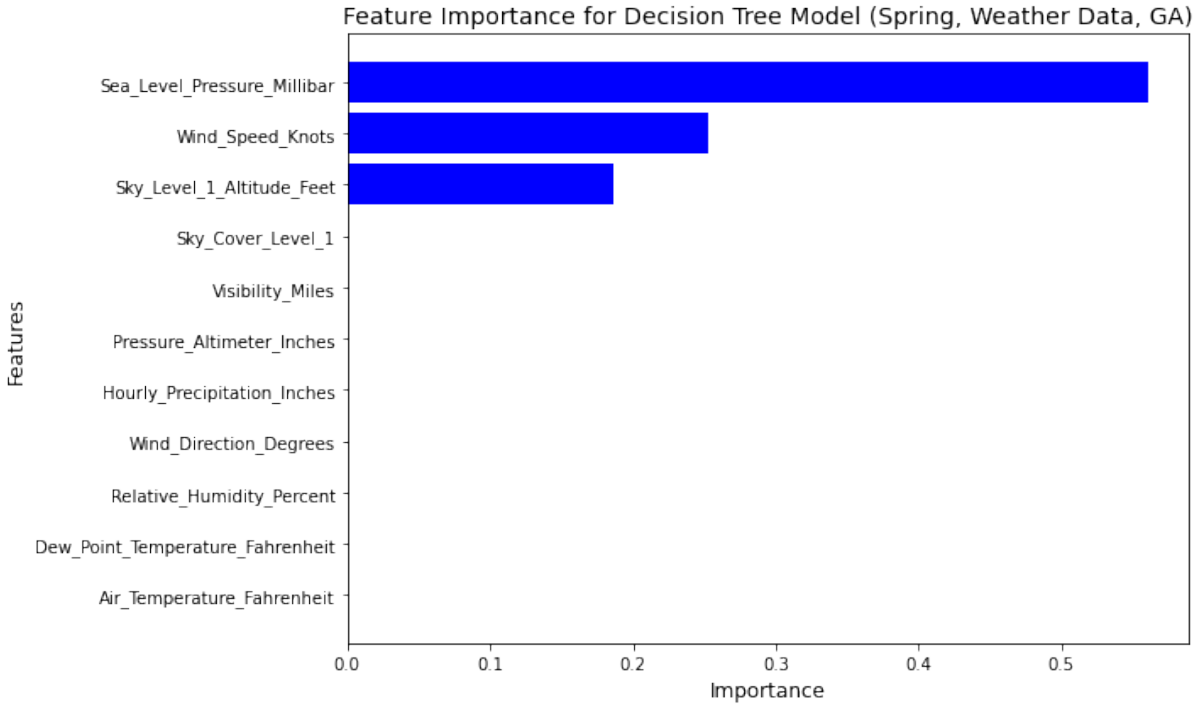


Figure 58: Feature Importance for January - Combined Weather and Flight Data

c) April

- Feature Importance for Georgia April - Combined Weather and Flight Data:

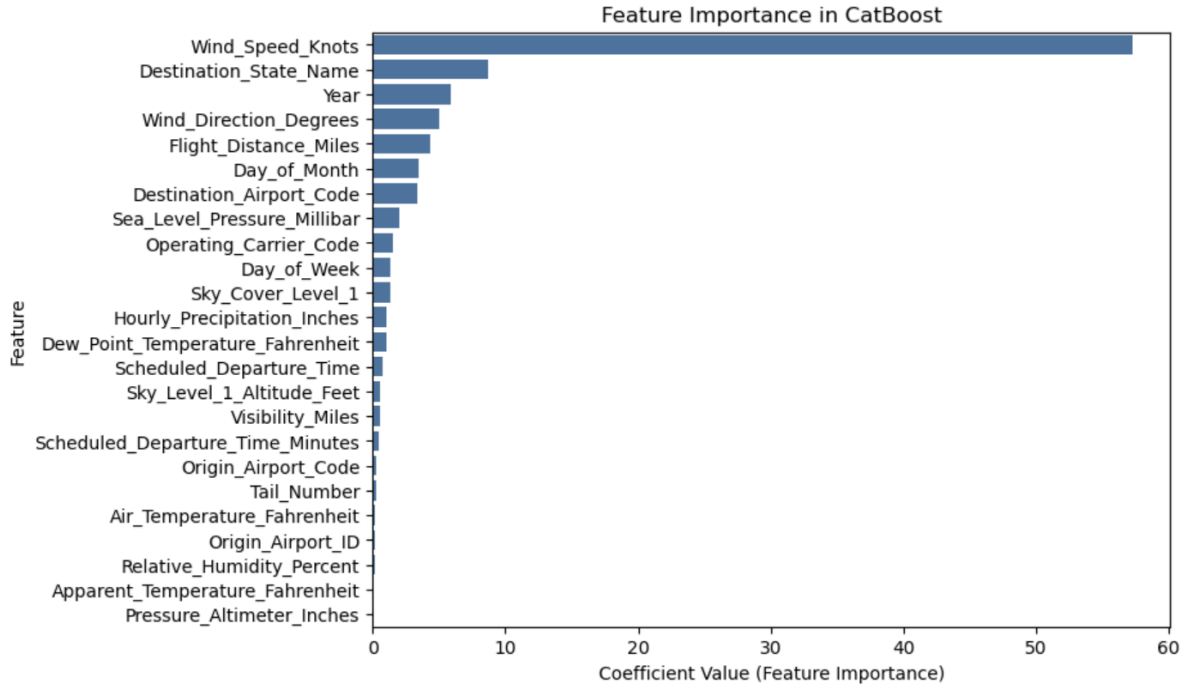


Figure 59: Feature Importance for April - Combined Weather and Flight Data

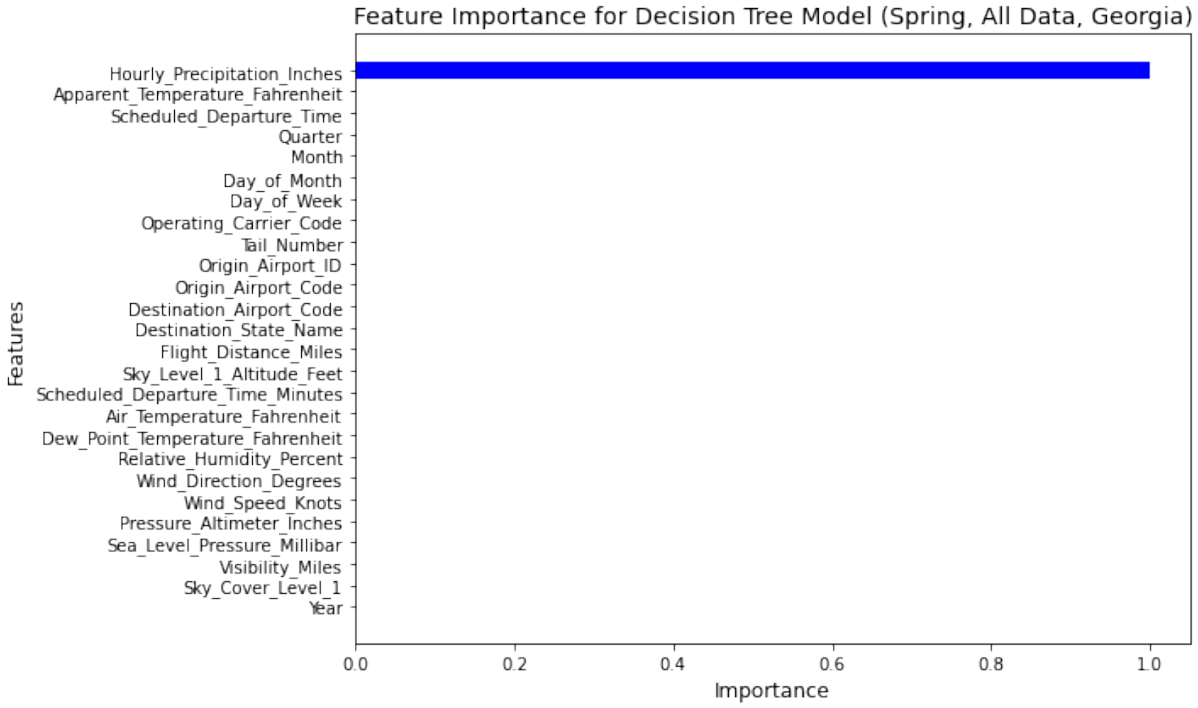


Figure 60: Feature Importance for April - Combined Weather and Flight Data

d) July

- Feature Importance for Georgia July - Combined Weather and Flight Data:

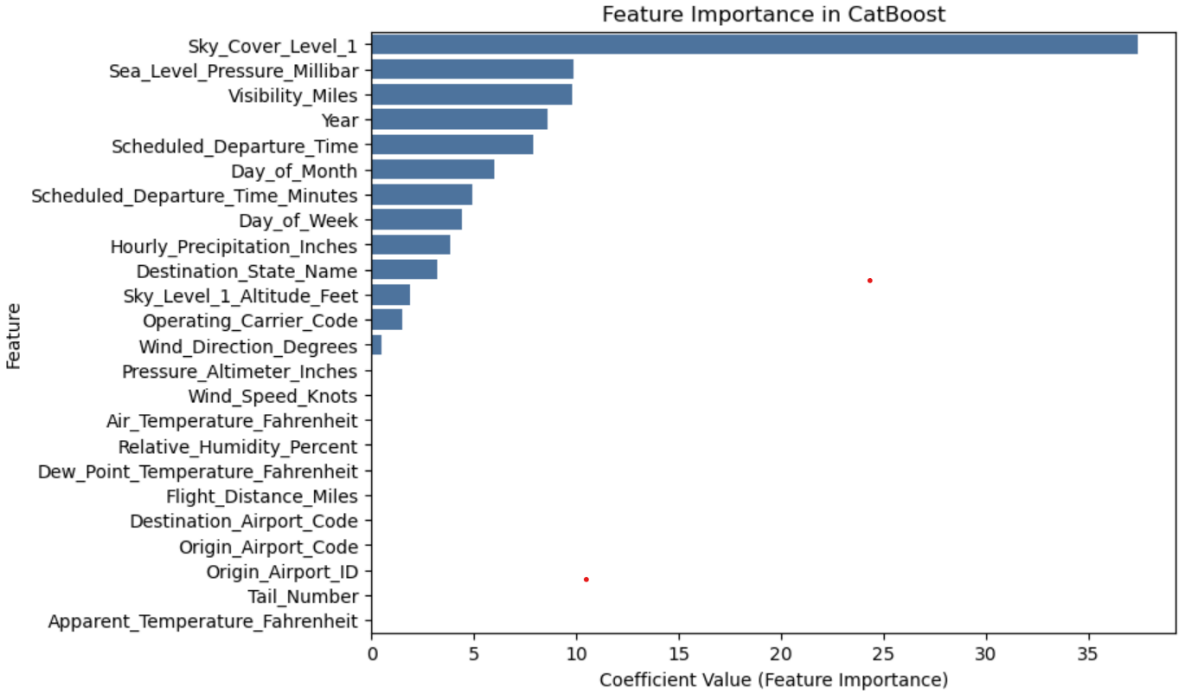


Figure 61: Feature Importance for July - Combined Weather and Flight Data

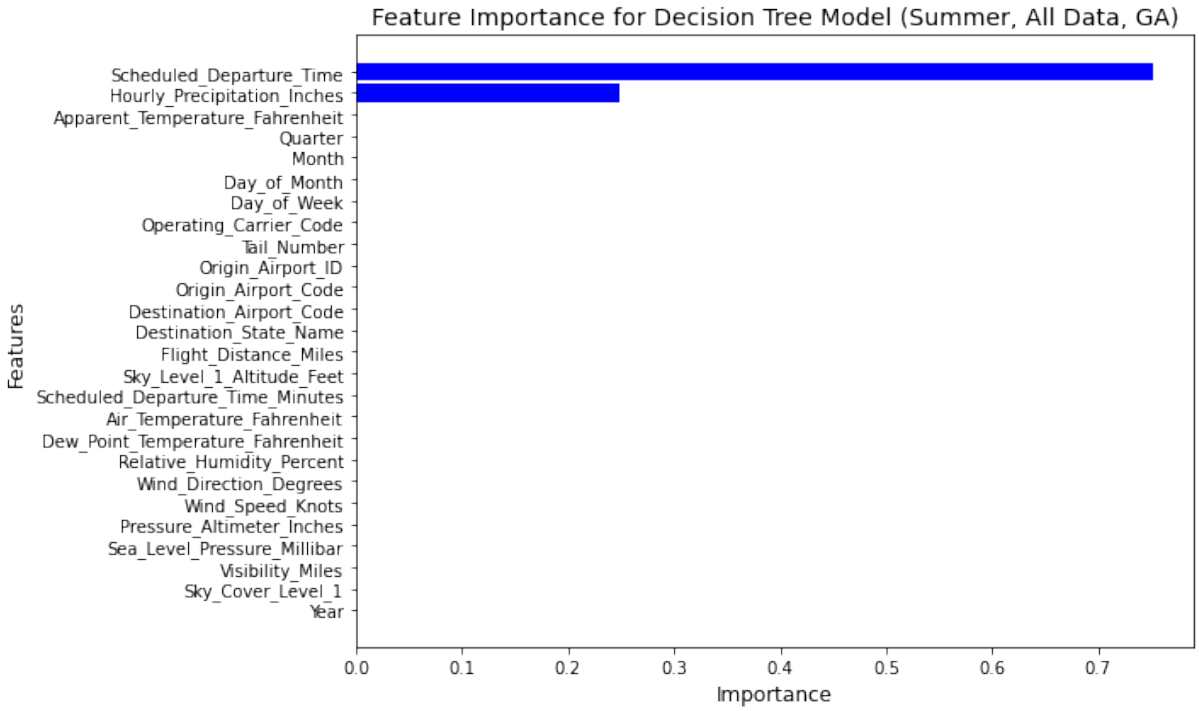


Figure 62: Feature Importance for July - Combined Weather and Flight Data

e) October

- Feature Importance for Georgia October - Combined Weather and Flight Data:

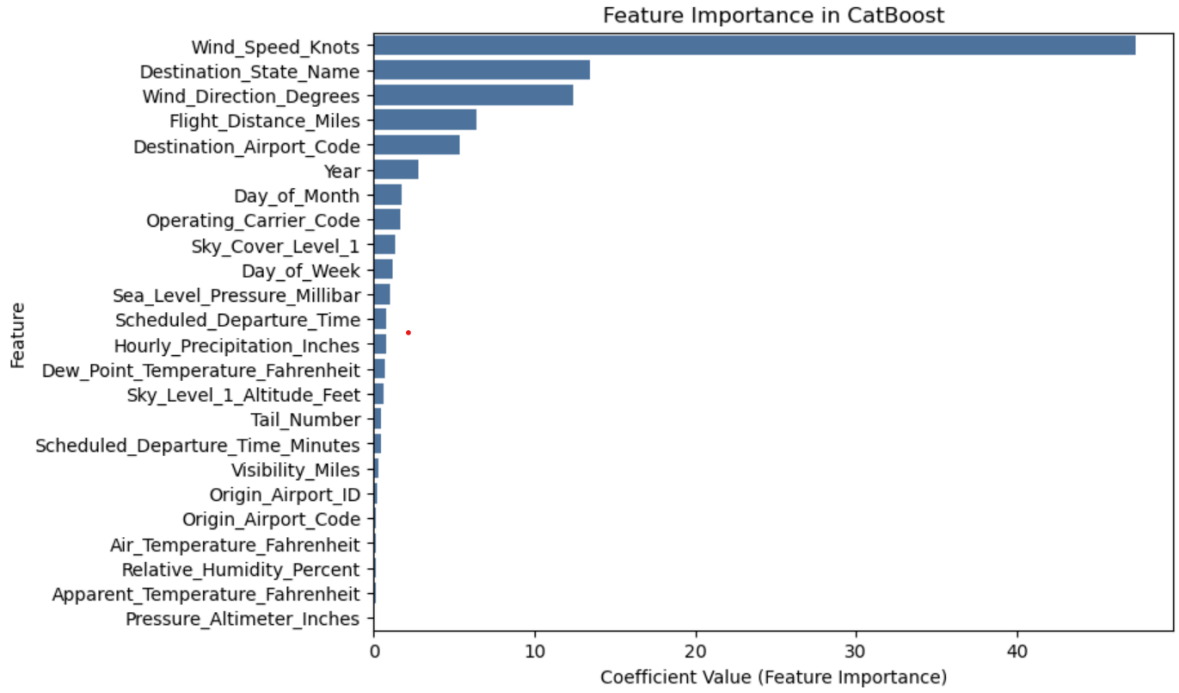


Figure 63: Feature Importance for October - Combined Weather and Flight Data

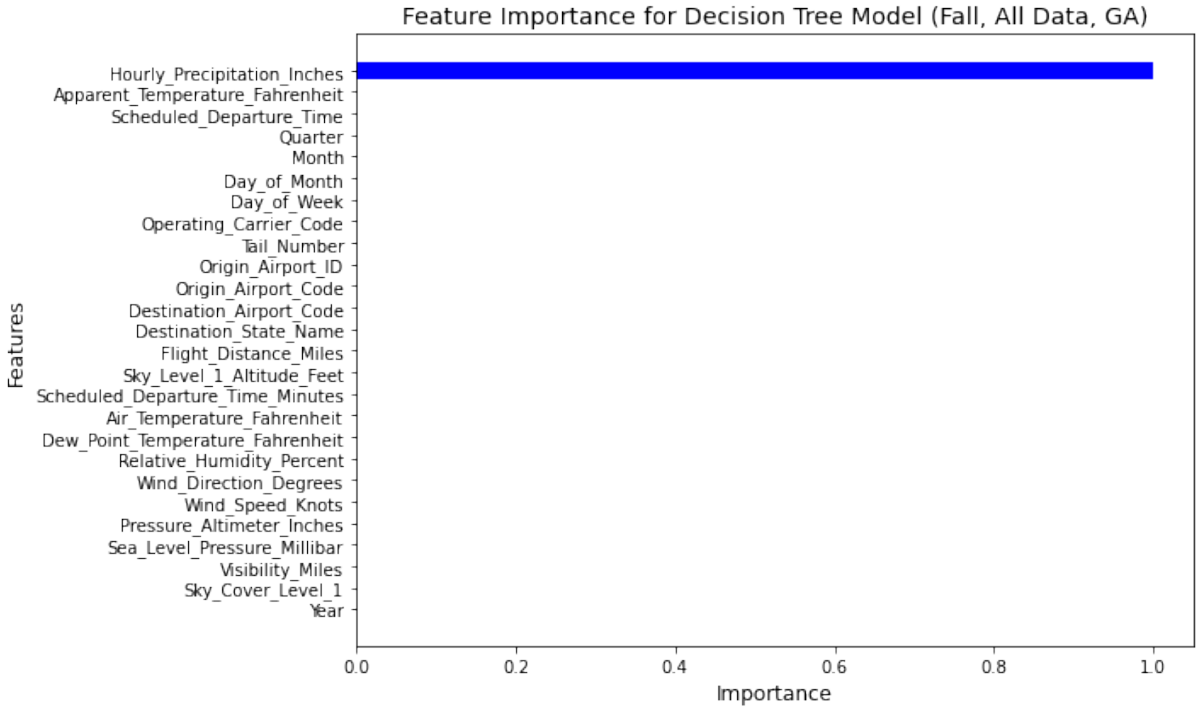


Figure 64: Feature Importance for October - Combined Weather and Flight Data

9 References

[1] Yhdego et al. "Analyzing the Impacts of Inbound Flight Delay Trends on Departure Delays Due to Connection Passengers Using a Hybrid RNN Model." Mathematics 2023, 11, 2427.

[2] Kim, S., Park, E. Prediction of flight departure delays caused by weather conditions adopting data-driven approaches. J Big Data 11, 11 (2024).

[3] Goodman, C. J., and J. D. Small Griswold, 2019: Meteorological Impacts on Commercial Aviation Delays and Cancellations in the Continental United States. J. Appl. Meteor. Climatol., 58, 479–494

[4] Kiliç, K.; Sallan, J.M. Study of Delay Prediction in the US Airport Network. Aerospace 2023, 10, 342

[5] Gratton, G. B. et al. "Reviewing the Impacts of Climate Change on Air Transport Operations." The Aeronautical Journal 126.1295 (2022): 209–221. Web.

[6] Tileagă, C., Oprișan, O. (2021). Flights Delay Compensation 261/2004: A Challenge for Airline Companies?. In: Orăștean, R., Ogrean, C., Mărginean, S.C. (eds) Organizations and Performance in a Complex World. IECS 2019. Springer Proceedings in Business and Economics. Springer, Cham.

[7] Carlier, Sandrine, Ivan de Lépinay, Jean-Claude Hustache, and Frank Jelinek. Environmental Impact of Air Traffic Flow Management Delays. CiteSeerX, doi:10.1.1.76.3545.

[8]Sekelová, I., Korba, P., Pjurová, S., Marimuthu, S., Kale, U. (2023). Reducing the Environmental Impact of Aviation by Minimizing Flight Delays. In: Sogut, M.Z., Karakoc, T.H., Secgin, O., Dalkiran, A. (eds) Proceedings of the 2022 International Symposium on Energy Management and Sustainability . ISEMAS 2022. Springer Proceedings in Energy. Springer, Cham.

[9] Hsu et al. "Unraveling Extreme Weather Impacts on Air Transportation and Passenger Delays using Location-based Data."

[10] "Prediction of Weather-induced Airline Delays Based on Machine Learning Algorithms." 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), Digital Avionics Systems Conference (DASC), 2016 IEEE/AIAA 35th.